

Actes de colloque / proceedings

4ème conférence internationale sur les Technologies de l'Information et de la Communication et l'Amazighe

ⵉⵙⴰⴳⴰⵍⵏ ⵉⵙⵏⵉⵍⵙⵏ : ⵜⵉⴳⴳⴰ ⴷ ⵓⵙⵎⵔⵙ

Les ressources langagières : Construction et Exploitation

Coordination
Ait ouguengay Youssef
Boulaknadel Siham

2012

LES RESSOURCES LANGAGIERES
CONSTRUCTION ET EXPLOITATION

LES RESSOURCES LANGAGIERES : CONSTRUCTION ET EXPLOITATION



Le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication (CEISIC)

Le Centre d'Aménagement Linguistique (CAL)

Le 4ème atelier international sur les Technologies d'Information et de Communication pour l'Amazighe

Sous le thème :

ⵉⵙⴰⴳⴰⵎⵏ ⵉⵙⵏⵉⵍⵙⵏ : ⵜⵉⵚⴽⴰ ⴷ ⵓⵙⵖⴰⵣⵓ

Les ressources langagières : Construction et Exploitation

Rabat, 24 – 25 février 2011

ⵇⵇⵙⴰⵟ, 24-25 ⵙⵔⴰⵖⵔ 2011

Publications de l'Institut Royal de la Culture Amazighe

Centre des Etudes Informatiques, des Systèmes d'Informations et de Communication

Série : Colloques et séminaires N° 27

Titre	:	Les ressources langagières : Construction et Exploitation
Coordination	:	Ait ouguengay Youssef et Boulaknadel Siham
Éditeur	:	Institut Royal de la Culture Amazighe
Réalisation et suivi	:	Centre des Etudes Informatiques, des Systèmes d'Information et de Communication.
Couverture	:	Unité d'édition - CTDEC
Dépôt légal	:	2012 MO 1393
ISBN	:	9954-28-117-8
Imprimerie	:	El Maarif Al Jadida
Copyright	:	® IRCAM

Conférence internationale sur : « Les ressources langagières : Construction et Exploitation »

ARGUMENTAIRE

Les ressources linguistiques interviennent, d'une façon croissante, au coeur même de la conception et du développement de différents produits informatiques, que ce soit en amont (utilisation de données linguistiques) ou en aval pour la production de nouvelles ressources et le développement de nouveaux outils.

Les langues naturelles peu informatisées souffrent en général d'un manque en termes de ressources, qui constituent un besoin crucial pour l'intégration de ces langues dans les nouvelles technologies d'information. Pour cette raison, une des préoccupations majeures du traitement automatique des langues (TAL) est la disposition de telles ressources. De nos jours, entamer un tel champ nécessite d'abord une préparation basique (codage de système graphique, claviers de saisie, etc.) de la langue en question avant de s'aligner sur les spécifications et normes internationales en la matière pour garantir un maximum de réutilisation des ressources et outils développés et d'interopérabilité avec les autres langues.

Dans le cas de la langue amazighe, le processus de standardisation et de généralisation de l'utilisation du caractère tifinaghe au niveau des technologies d'information, a préparé le terrain pour le développement des outils du TAL et pour la gestion des ressources linguistiques, aussi bien monolingues que multilingues. Il reste néanmoins d'autres efforts à consentir pour favoriser l'usage de l'amazighe et contribuer à sa promotion.

PREFACE

La sélection d'articles publiés dans le présent recueil constitue les actes du 4ème atelier international sur l'amazighe et les technologies d'information et de communication (TIC) qui s'est tenu à l'IRCAM du 24 au 25 février 2011.

L'objectif de ce 4ème atelier est de donner une vue sur les efforts des différents chercheurs nationaux et internationaux travaillant en traitement automatique des langues naturelles en particulier l'amazighe et de renforcer la culture de la mutualisation et le partage des ressources langagières.

Les travaux réunis dans ce recueil traduisent à la fois le caractère multidisciplinaire des recherches, la richesse des applications sous-jacentes et la vitalité des innovations issues du traitement automatique des langues.

Lors de cette quatrième édition, sur les 31 soumissions reçues, 24 articles ont été sélectionnés par un comité de lecture important. En général deux relecteurs ont été mis à contribution pour chaque article.

Nos remerciements chaleureux vont tout d'abord aux auteurs pour leurs contributions scientifiques. Nous remercions également les membres du comité de lecture, pour leurs rapports d'évaluation précis et constructifs et le temps qu'ils y ont consacré.

Nos vifs remerciements vont également à toute l'équipe du Comité d'organisation pour leur mobilisation permanente, leur travail, et leur enthousiasme communicatif pour faire du 4ème atelier international sur l'amazighe et les TICs une grande réussite. Qu'ils sachent que nous avons été nombreux à avoir été touchés par leurs attentions et leurs actions.

Nous espérons que de nombreux chercheurs et experts, intéréssés par l'amazighe et voulants explorer ce domaine, trouverons cet ouvrage utile.
Les étudiants chercheurs pourront y trouver aussi une aide précieuse quant aux questions sur les approches et applications du TAL.

Table des matières

Vers l'informatisation de quelques langues d'Afrique de l'Ouest

Chantal Enguehard
chantal.enguehard@univ-nantes.fr
Laboratoire d'Informatique de Nantes Atlantique
France

Soumana Kané
soumanak@yahoo.com
Centre National des Ressources de l'Éducation Non Formelle
Mali

Mathieu Mangeot
Mathieu.Mangeot@imag.fr
Laboratoire d'Informatique de Grenoble
France

Issouf Modi
modyissouf@yahoo.fr
Ministère de l'Éducation Nationale
Direction Générale de l'enseignement de base
Niger

Mamadou Lamine Sanogo
mala_sng@yahoo.fr
Centre National de la Recherche Scientifique et Technologique
Burkina Faso

Si l'accès aux ordinateurs est considéré comme le principal indicateur de la fracture numérique en Afrique, il faut reconnaître que la disponibilité des ressources dans les langues africaines constitue un handicap dont les conséquences sont incalculables pour le développement des Technologies de l'Information et de la Communication (TIC) dans cette partie du monde. Aussi, la production, la diffusion et la vulgarisation de ressources locales adaptées dans ces langues nous paraissent-elles être indiquées pour une implantation durable des TIC sur le continent. Or, la plupart des langues de l'espace francophone d'Afrique de l'Ouest sont peu dotées (langues-pi) [Berment 2004] : les ressources électroniques disponibles sont rares, mal distribuées, voire inexistantes. Seules sont accessibles les fonctions d'édition et d'impression rendant l'exploitation de ces langues difficile au moment où il est question de les introduire dans le système éducatif, de créer des normes d'écriture standardisées et stabilisées et surtout de développer leur usage à l'écrit dans l'administration et la vie quotidienne.

Aussi, afin de contribuer à combler ce retard, nous nous sommes engagés -avec les collègues du Sud et du Nord- à améliorer l'équipement de quelques langues africaines à travers, entre autres, l'informatisation de dictionnaires éditoriaux portant sur des langues africaines. A cet effet, nous présenterons le projet DiLAF (Dictionnaires Langues Africaines Français) qui vise à convertir des dictionnaires

éditoriaux bilingues en un format XML[1] permettant leur pérennisation et leur partage [Streiter et al. 2006]. Ce projet international rassemble des partenaires du Burkina Faso (Centre National de la Recherche Scientifique et Technologique), de France (Laboratoire d'Informatique de Grenoble et Laboratoire d'informatique de Nantes-Atlantique), du Mali (Centre National de Ressources de l'Éducation Non Formelle) et du Niger (Institut National de Documentation de Recherche et d'Animation Pédagogiques, Ministère de l'Education Nationale, et Université Abdou Moumouni de Niamey).

En nous fondant sur un travail de base déjà effectué par des lexicographes nous avons constitué des équipes pluridisciplinaires constituées de linguistes, d'informaticiens et de pédagogues. Cinq dictionnaires comportant, chacun, plusieurs milliers d'entrées, devraient être convertis et intégrés à une plate-forme Jibiki de gestion de ressources lexicales [Mangeot 2001]. Les dictionnaires seront donc disponibles sur Internet d'ici la fin de l'année 2011 sous licence Creative Commons.

— dictionnaire bambara-français, Charles Bailleul, édition 1996,
— dictionnaire haoussa-français destiné à l'enseignement du cycle de base 1, 2008, Soutéba,
— dictionnaire kanouri-français destiné pour le cycle de base 1, 2004, Soutéba,
— dictionnaire soŋay zarma-français destiné pour le cycle de base 1, 2007, Soutéba,
— dictionnaire tamajaq-français destiné à l'enseignement du cycle de base 1, 2007, Soutéba.

Il s'agit de dictionnaires d'usage qui visent surtout à vulgariser les formes écrites de l'usage quotidien des langues africaines dans la pure tradition lexicographique [Matoré 1973], [Eluerd 2000].

Se démarquant des démarches normatives et dirigistes des dictionnaires normatifs [Mortureux 1997], les présents dictionnaires descriptifs restent ouverts aux contributions et leur mise en ligne devra, nous l'espérons, développer un sentiment de fierté chez les usagers des différentes langues. De même, ils participeront au développement d'un environnement lettré propice à l'alphabétisation dont le faible taux compromet les acquis des progrès réalisés dans les autres secteurs.

Nous présenterons la structure de ces dictionnaires ainsi que quelques entrées, puis

---

[1]Extended Markup Language

contributions et leur mise en ligne devra, nous l'espérons, développer un sentiment de fierté chez les usagers des différentes langues. De même, ils participeront au développement d'un environnement lettré propice à l'alphabétisation dont le faible taux compromet les acquis des progrès réalisés dans les autres secteurs.

Nous présenterons la structure de ces dictionnaires ainsi que quelques entrées, puis les résultats de l'atelier de démarrage qui s'est déroulé du 6 au 17 décembre 2010 à Niamey (Niger) :

— méthodologie de conversion à Unicode,
— formation aux expressions régulières,
— méthodologie de conversion à XML.

Nous présentons l'origine des dictionnaires, quelques entrées ainsi que leur structure puis, nous détaillons les premiers résultats de l'atelier tout en nous projetant vers les futurs travaux.

1. Cinq dictionnaires bilingues langue africaine-français

Quatre des cinq dictionnaires sur lesquels nous travaillons ont été produits par le projet Soutéba (programme de soutien à l'éducation de base) avec le financement de la coopération allemande[2] et l'appui de l'Union Européenne. Ces dictionnaires, destinés à l'éducation de base, sont de structure simple car ils ont été conçus pour des enfants de classe primaire scolarisés en école bilingue (l'enseignement y est donné en une langue nationale et en français). La plupart des termes de lexicologie, telles les étiquettes lexicales ou les catégories grammaticales, les signalisations de synonymies, d'antonymies, de genres, de variations dialectales, etc., y sont notés dans la langue dont il est question dans le dictionnaire, contribuant ainsi à forger et à diffuser un méta-langage dans la langue locale, une terminologie spécialisée. Les entrées sont énoncées en ordre alphabétique, même dans le cas du tamajaq (bien qu'il soit habituel de présenter les entrées de cette langue en fonction des racines) car les voyelles sont explicitement écrites (ce mode de classement a été privilégié car il est bien connu des enfants).

[2] DED : Deutscher Entwicklungsdienst

1.1 - Dictionnaire haoussa-français

Il comprend 7823 entrées classées selon l'ordre lexicographique suivant : a b ɓ c d e f fy g gw gy h i j k kw ky ƙ ƙw ƙy l m n o p r s sh t ts u w y ƴ z [Arrêté 212-99].

Elles sont structurées avec des schémas différents selon la catégorie grammaticale. Toutes les entrées sont d'ordre orthographique ; suivent la prononciation (les tons sont marqués par les signes diacritiques posés sur les voyelles) et la catégorie grammaticale. Sur le plan sémantique, il existe une définition en langue haoussa, un exemple d'emploi (repéré par l'usage de l'italique), puis l'équivalent en français. L'entrée d'un nom précise en sus le genre, le féminin s'il existe, le ou les pluriels (selon les genres) et les éventuelles variantes dialectales. Pour les verbes, il est parfois nécessaire de préciser les degrés pour calculer les dérivés morphologiques. Les variantes morpho- phonologiques des dérivations féminine et plurielle des adjectifs sont énoncées.

Exemple :

> **jaki** [jàakíi] *s*. **babbar dabbar gida mai kamar doki, wadda ba ta kai tsawon doki ba**
>
> **amma ta fi shi dogayen kunnuwa**. ***Ya aza wa jaki kaya za ya tafi kasuwa.*** *Jin*.: n. *Sg*.: **jaka**.

La forme orthographique de l'entrée est suivie d'indication de prononciation ciblées sur la notation des tons : le ton haut est noté par un accent aigu, le ton bas par un accent grave, le ton montant par un caron (signe suggérant la succession d'un accent grave et d'un accent aigu) et le ton descendant par un accent circonflexe (signe suggérant la succession d'un aigu et d'un accent grave. La catégorie grammaticale de l'entrée est indiquée en italique. Une définition, un exemple d'usage puis le sens en français sont ensuite énoncés. D'autres informations peuvent apparaître comme des variantes.

Exemple :

> **abəɼwa** [àbəaɼwà] *cu*. **Kəska təngər7i, kalu ngəwua dawulan tada cakkidə.** ***Kəryende***
>
> ***kannua nangaro, abərwa cakkiwawo.*** [*Fa*.: ananas]
>
> Le mot "abəɼwa" se prononce [àbə□wà]. Sa catégorie grammaticale est "cu." (nom).
>
> Sa définition est écrite en caractères gras : "Kəska təngərmi, kalu ngəwua dawulan tada
>
> cakkidə."

Un exemple d'usage est signalé en caractères italique : "Kəryende kannua nangaro, abərmwa

cakkiwawo."

L'équivalent en français, précédé de "Fa.:" et encadré de crochets, termine l'entrée.

1.3 - Dictionnaire soŋay zarma-français

Il comprend 6916 entrées classées selon l'ordre lexicographique suivant :

a ã b c d e e f g h i ĩ j k l m

n ŋ ɲ o õ p r s t u ũ w y z [Arrêté 215-99].

Chaque entrée présente une forme orthographique suivie d'une transcription phonétique dans laquelle les tons sont notés selon les conventions déjà exposées pour le kanouri (partie 1.2). La catégorie grammaticale précise explicitement, pour les verbes, la transitivité ou l'intransitivité. Pour certaines entrées, des antonymes, synonymes ou renvois sont indiqués. Une glose en français, une définition et un

exemple terminent l'entrée.

Exemple :

ɲ agas [ɲ agas] *mteeb.* ● *brusquement (détaler)* ● sanniize no kaŋ ga cabe kaŋ boro na zuray

sambu nda gaabi sahã-din ● *Za zankey di hansu-kaaro no i te ɲ agas*

Le mot "**ɲ agas**" se prononce [ɲ agas]. Sa catégorie grammaticale est "*mteeb.*" (adverbe). L'équivalent en français est signalé en caractères italiques.

Sa définition est : "sanniize no kaŋ ga cabe kaŋ boro na zuray sambu nda gaabi sahã-din"

Un exemple d'usage est énoncé en caractères italiques : "*Za zankey di hansu-kaaro no i te*

ɲ agas"

1.4 - Dictionnaire tamajaq-français

Le dictionnaire tamajaq-français comprend 5205 entrées du parler təwəlləmmət classées selon l'ordre lexicographique suivant : a â ă ə b c d ḍ e ê f g ğ h i î j ǰ ɣ k l ḷ m n ŋ o ô q r s ṣ š t ṭ u û w x y z ẓ [Arrêté 214-99] [3].

La forme orthographique de l'entrée est suivie de la catégorie grammaticale de l'entrée et d'une glose en français indiquées en italique. Pour les noms figurent souvent des indications morphologiques concernant l'état d'annexion ; le pluriel et le genre sont souvent explicitement indiqués. Une définition, un exemple d'usage sont ensuite énoncés. D'autres informations peuvent apparaître comme des variantes, des synonymes, etc.. Le tamajaq n'étant pas une langue tonale, la phonétique n'apparaît pas.

[3] Les signes 'ǰ' et ' ğ' sont utilisés uniquement pour transcrire certains parlers comme celui de l'Ayər, par conséquent ils

n'apparaissent pas dans ce dictionnaire.

Exemple :

əbeɣ la *sn. mulet* ♦ **Ag-anɣ er əd tabagawt. Ibeɣ lan wər tan-taha tamalaɣ a.** *anammelu.:*
fakr-ejaḍ. *təmust.:* **yy.** *iget.:* **ibəɣ lan.**

Le mot "**əbeɣ la**" est un "sn.", abbréviation de "isən" (nom) qui signifie mulet en français.

Sa définition "Ag-anɣ er əd tabagawt." et un exemple d'usage "Ibeɣ lan wər tan-taha tamalaɣ " sont écrits en caractères gras.

Un synonyme (anammelu) est signalé : "fakr-ejaḍ".

Le genre (təmust) est "yy.", abréviation de "yey" (masculin).

Le pluriel de ce mot (iget) est "ibəɣ lan".

1.5 - Dictionnaire bambara-français

Le dictionnaire bambara-français du Père Charles Bailleul (édition 1996) comprend plus de 10 000 entrées selon l'ordre lexicographique suivant : a b c d e ɛ f g h i j k l m n ɲ ŋ o ɔ p r s t u w y z.

Ce dictionnaire est d'abord destiné aux locuteurs français désireux de se perfectionner en bambara mais il constitue également une ressource pour les bambaraphones. Selon les dires de l'auteur lui- même, il « se veut être un outil de travail au service de l'alphabétisation, l'enseignement et la culture bambara ». A ce jour, il peut être considéré comme le dictionnaire le plus fourni et le plus complet sur cette langue. Aussi il est consulté par les spécialistes des autres variétés de cette langue que sont le dioula (Burkina Faso, Côte d'Ivoire) et le manlinké (Guinée, Gambie, Sierra Leone, Libéria, etc.).

Bien que l'orthographe du bambara ne note pas les tons, et ce par économie de signes, les tons sont marquées dans toutes les entrées et tous les exemples d'usage : l'accent grave sur une voyelle brève marque un ton bas ponctuel ("**bìnɔ gɔ kɛ** " – "oncle paternel") ; l'accent grave sur une voyelle répétée l'affecte sur toute sa longueur ("**dɛ Wɛ mu**" – "parole" – se prononce **dɛ Wɛ Wmu**); l'accent grave suivi d'un accent aigu marque une voyelle longue relevée sur sa deuxième

partie (ex : "**ɲ àá**" – "nid)" ; le caron marque un ton bas modulé ascendant (ex : "**ben**" – "accord").

La prononciation phonétique n'est indiquée que lorsque l'orthographe officielle s'écarte de la prononciation effective. Dans de tels cas, elle est indiquée entre crochets. Par exemple L'analyse de « da.lan [dlan] (se coucher.suff instrument) n. lit » montre que ce dérivé ("da" et le suffixe "-lan", respectivement "se coucher" et "instrument servant à"" n'est jamais prononcé complètement c'est-à- dire en deux syllabes, il est phonétiquement noté par [dlan].

Les entrées, surtout complexes, sont accompagnées de leur origine et de leur structure, car il s'agit d'informations nécessaires pour une bonne traduction. Ainsi, pour les dérivés et composés, l'analyse des éléments est indiquée entre parenthèses et la frontière sémantique suggérée par un point, comme dans l'entrée suivante : « **ɲ ɛ mɔ gɔ** ɲ ɛ .mɔ gɔ (devant.personne) dirigeant, chef. [...] » Cette présentation de l'entrée indique que, morphologiquement, "**ɲ ɛ mɔ gɔ** " se compose de "**ɲ ɛ** " et de "**mɔ gɔ** " (ce qui est indiqué par le point) et que, sémantiquement, dans l'ordre, il signifie "**devant**" et "**personne**" (ce qui est indiqué par les parenthèses et le point), le sens de tout le composé se ramenant à dirigeant, c'est-à-dire une personne placée devant, à la tête de... (traduction indiquée par le soulignement).

On peut ainsi multiplier les exemples :

« **kalanso** kàlàn.so (instruction.maison) classe d'école » : mot composé de "kalan" et "so",

respectivement "instruction" et "maison", signifie "classe d'école".

« **mɔ gɔ dun** mɔ gɔ .dun (personne.manger) cannibale, anthropophage » : mot composé de "mɔ gɔ " et "dun", respectivement "personne" et "manger", signifie "cannibale".

« **juguya** jugu.ya (mauvais.suff abst) méchanceté » : mot dérivé ("jugu" et "-ya", respectivement

"mauvais" et suffixe d'abstraction), signifie "méchanceté".

« **walanba** walan.ba (tablette.suff augm) tableau noir » : mot dérivé ("walan" et "-ba",

respectivement "tablette" et suffixe augmentatif), signifie "tableau noir".

Il est important de signaler que la dérivation et la composition étant des procédés

très productifs en bambara, les cas retenus pour figurer dans le dictionnaire ont été choisis en fonction de leur fréquence d'emploi et de leur variation de sens par rapport à leur formation.

L'origine des emprunts est indiquée entre accolades : {fr} pour le français, et {ar} pour l'arabe.

Exemples : « **kaso** kàso {fr: cachot} n. Prison » ; « **ala** ala {ar: allah=Dieu} »

Enfin, ce dictionnaire accorde quelque place aux néologismes proposés par les services d'alphabétisation. Il s'agit notamment de « ceux qui sont les plus utilisés ou semblent promis à un bel avenir ». Ils sont signalés par l'indication (néologisme).

Exemples : « **kumaden** kuma.den (parole.élément) mot (néologisme) » ; « **kɔ bila** kɔ .bila (derrière.placer) postposition (néologisme) »

2. Plate-forme jibiki

Jibiki (Mangeot et al., 2003; Mangeot et al., 2006) est une plate-forme générique en ligne pour manipuler des ressources lexicales avec gestion d'utilisateurs et groupes, consultation de ressources hétérogènes et édition générique d'articles de dictionnaires. C'est un site Web communautaire initialement développé pour le projet Papillon (http://www.papillon-dictionary.org). La plate-forme est programmée entièrement en Java, fondée sur l'environnement "Enhydra". Toutes les données sont stockées au format XML dans une base de données (Postgres). Ce site Web propose principalement deux services : une interface unifiée permettant d'accéder simultanément à de nombreuses ressources hétérogènes (monolingues, dictionnaires bilingues. bases multilingues, etc.) et une interface d'édition spécifique pour contribuer directement aux dictionnaires disponibles sur la plate-forme.

L'éditeur (Mangeot et al., 2004) est fondé sur un modèle d'interface HTML instancié avec l'article à éditer. Le modèle peut être généré automatiquement depuis une description de la structure de l'entrée à l'aide d'un schéma XML. Il peut être modifié ensuite pour améliorer le rendu à l'écran. La seule information nécessaire à l'édition d'un article de dictionnaire est donc le schéma XML représentant la structure de cette entrée. Par conséquent, il est possible d'éditer n'importe quel type de dictionnaire s'il est encodé en XML.

Plusieurs projets de construction de ressources lexicales ont utilisé ou utilisent toujours cette plate- forme avec succès. C'est le cas par exemple du projet GDEF (Chalvin et al., 2006) de dictionnaire bilingue estonien-français (http://estfra.ee), du projet LexALP de terminologie multilingue sur la convention alpine (http://lexalp.eurac.edu/) ou plus récemment du projet MotÀMot sur les langues d'Asie du sud-est. Le code de cette plate-forme est disponible gratuitement en source ouverte en téléchargement depuis la forge du laboratoire LIG (http://jibiki.ligforge.imag.fr).

La plate-forme sera adaptée spécifiquement au projet DiLAF car, en sus des dictionnaires, des informations spécifiques au projet doivent être accessibles aux visiteurs :

— présentation du projet et des partenaires

— méthodologie générale de conversion des dictionnaires éditoriaux au format LMF

— fiches techniques concernant différents outils ou tâches à réaliser : tutoriel sur les expressions régulières, méthodologie de conversion d'un document utilisant des polices non conformes au standard Unicode vers un document conforme au standard Unicode, liste des logiciels utilisés (il s'agit uniquement de logiciels libres), méthodologie de suivi du projet.

— présentation de chaque dictionnaire : genèse, auteurs initiaux, principes ayant régi la construction du dictionnaire, langue, alphabet, structuration des articles, etc.

— dictionnaire au format LMF.

Il est également envisagé de localiser la plate-forme pour chacune des langues du projet en traduisant les libellés de l'interface.

DiLAF

ɲagas [ɲágás] **mteeb.** brusquement (dételer)

sanniize no kaŋ ga cabe kaŋ boro na zuray sambu nda gaabi sahã-din

Za zankey di hansu-kaaro no i te ɲagas

Figure 1 : présentation du verbe zarma "ɲ agas" sur la plate-forme jibiki

3. Travaux du premier atelier du projet DiLAF

Les participants à cet atelier sont majoritairement des linguistes ou des pédagogues, chacun travaillant sur un dictionnaire traitant de sa langue maternelle (qui est également la langue sur laquelle portent ses activités professionnelles). Les formateurs sont des enseignants-chercheurs en informatique spécialisés en traitement automatique des langues (TAL). L'objectif de ce premier atelier est de délivrer une formation à la conversion des dictionnaires tels qu'ils existent dans leur format éditorial, vers une structure XML reflétant au mieux la structure initiale des entrées tout en conservant l'ensemble des informations qui y sont exprimées. Plusieurs étapes ont été suivies pour atteindre cet objectif et garder la trace des différents traitements, chacune de ces étapes étant assortie d'un document remis aux participants.

Formation aux expressions régulières

Les participants ont été formés à l'usage des expressions régulières pendant trois jours et ont pu exercer directement leurs nouvelles connaissances par l'usage du logiciel Open Office Writer.

Conversion à Unicode

Bien que les alphabets des langues sur lesquels nous avons travaillé soient majoritairement d'origine latine, de nouveaux caractères nécessaires pour noter des sons spécifiques à certaines langues[4] à l'aide d'un seul caractère[5] ont été adoptés par les linguistes lors d'une série de réunions[6]. La première, en septembre 1978, organisée par l'UNESCO au CELTHO (Centre d'études linguistiques et historiques par tradition orale) à Niamey crée l'« Alphabet africain de référence » fondé sur les conventions de l'IPA (International Phonetic Association) et de l'IAI (International African Institute). Ainsi, chacun des alphabets que nous avons précédemment

[4] L'absence d'un seul signe marquant certains sons avait amené les linguistes africains à exprimer ces sons à l'aide de combinaisons de lettres. Par exemple, en zarma le digraphe /ny/ note le son n palatal. C'est aussi ce qui est réalisé en français avec le son [ʃ] retranscrit /ch/.

[5] En Zarma, la lettre ɲ remplace le digraphe /ny/. Ainsi, le mot autrefois écrit « nya » (mère) devient « ɲ a ».

[6] Niamey (novembre 1978), Abidjan (décembre 1980), Bamako (juin 1981), Nouakchott (novembre 1981), Ouagadougou (juin 1982).

présentés comprend au moins un de ces "nouveaux" caractères : ɓ ɗ ə ɛ ʏ ƙ ɲ ŋ ɔ ʏ. Des caractères composés d'un caractère latin et d'un signe diacritique ont également été créés : â ê î ô û ă ã e ĩ õ ũ ḍ ḷ ṣ ṭ ẓ ğ ǰ š ɍ.

Comme nombre de ces caractères étaient absents des dispositifs de saisie et des standards alors en usage [Enguehard 2009], des touches de frappe de machines à écrire, des glyphes de polices d'ordinateurs ont été modifiées. Bien que la plupart de ces caractères soient depuis plusieurs années présents dans le standard Unicode (issu des travaux du comité ISO 10646 [Haralambous 2004]), les dictionnaires dont nous disposons ont été rédigés en utilisant les anciennes polices arrangées.

Une méthodologie a été définie afin de repérer et remplacer les caractères inadéquats par les caractères définis dans le standard Unicode. Suivre cette méthodologie implique que l'ensemble des caractères repérés et leurs caractères de remplacement soient notés dans un fichier afin de pouvoir réitérer facilement cette opération si cela s'avérait nécessaire.

Ce travail est terminé et a permis de dresser la liste des caractères encore absents d'Unicode ou dont la manipulation peut poser des problèmes avec certains logiciels (voir partie 4).

Méthodologie de conversion à XML

Les fichiers électroniques des dictionnaires respectant le standard Unicode ont été convertis en fichier Open Office. Ces fichiers sont en réalité des fichiers XML compressés, les balises exprimant principalement des informations relatives à la mise en forme (usage de caractères gras ou italiques, de couleur, etc.). Il s'agit donc de passer d'un format XML dédié à l'expression de la forme vers un format XML porteur d'informations sur la structure du dictionnaire : vedette, phonétique, exemple, synonymes, etc.

Cette transformation a été partiellement ou totalement réalisée à l'aide d'expressions régulières.

4. Bilan quant à Unicode

Certains caractères des alphabets sur lesquels nous avons travaillé nécessitent d'apparaître dans le standard Unicode ou d'être mieux pris en compte par les logiciels existants.

Ordre lexicographique des digraphes

Les digraphes peuvent être facilement composés à l'aide de deux caractères mais leur usage modifie l'ordre du tri lexicographique qui conditionne la présentation des entrées du dictionnaire. Ainsi, en haoussa et en kanouri, le digraphe 'sh' est situé après la lettre 's'. Donc le verbe "sha" (boire) est situé après le mot "suya" (frite) dans le dictionnaire haoussa, et le verbe"suwuttu" (dénouer) précède le nom "sһadda" (basin) en kanouri.

Ces subtilités peuvent être difficilement traitées au niveau logiciel et nécessiterait que les digraphes apparaissent en tant que signe dans le répertoire Unicode. Certains, utilisés par d'autres langues, y figurent déjà, parfois sous leur différentes casses : 'DZ' (U+01F1), 'Dz' (U+01F2), 'dz' (U+01F3) sont utilisés en slovaque ; 'NJ' (U+01CA), 'Nj' (U+01CB), 'nj' (U+01CC) en croate et pour transcrire la lettre « Њ » de l'alphabet cyrillique en serbe ; etc.

Il serait nécessaire de compléter le standard Unicode avec les digraphes des alphabets kanouri et haoussa sous leurs différentes casses.

fy	gw	gy	ky	kw	ƙy	ƙw	sh	ts
Fy	Gw	Gy	Ky	Kw	Ƙy	Ƙw	Sh	Ts
FY	GW	GY	KY	KW	ƘY	ƘW	SH	TS

Table 1 : digraphes du haoussa et du kanouri absents de Unicode

Caractères avec signes diacritiques

Certains des caractères potant des signes diacritiques figurent dans une Unicode comme un unique signe, d'autres ne peuvent être obtenus que par composition.
Ainsi, les voyelles avec tilde 'a', 'i', 'o' et 'u' figurent dans Unicode sous leurs formes minuscule et majuscule[7] tandis que le 'e' avec tilde est absent et doit être composé à l'aide du caractère 'e' ou 'E' suivi de l'accent tilde (U+303), ce qui peut provoquer des rendus différents des autres lettres avec tilde lors de l'affichage ou de l'impression (tilde situé à une hauteur différente par exemple).

La lettre j avec caron existe dans Unicode en tant que signe ǰ (U+1F0), mais sa forme majuscule doit être composée J̌ avec la lettre J et le signe caron (U+30C).
Les caractères e, E et J³ devraient être ajoutés au standard Unicode.

[7] 'ã' (U+00E3) 'ĩ '(U+0129), 'õ' (U+00F5), 'ũ' (U+O169), 'Ã' (U+00C3), 'Ĩ' (U+0128), 'Õ' (U+00D5) et 'Ũ' (U+0168).

Editeurs de texte : fonctions changement de casse, affichage et rechercher

Les éditeurs de texte disposent généralement de la fonction changement de casse, mais ne la réalisent pas toujours de manière correcte selon les caractères. Ainsi, nous avons constaté durant nos travaux que le logiciel OpenOffice Writer (version 3.2.1) échoue dans la transformation de 'ɍ' en ' Ɍ' du bas de casse vers le haut de casse ou pour l'inverse (le caractère reste inchangé) tandis que Notepad++ (version 5.8.6) échoue dans la transformation de ǰ en J[3] du bas de casse vers le haut de casse ou pour l'inverse (le caractère reste inchangé).

Plusieurs caractères avec diacritiques peuvent être directement saisis comme un seul signe (quand celui-ci existe dans Unicode) ou être explicitement composés. Selon les logiciels, les différentes versions d'un même caractère avec diacritiques peuvent être traités de manière égale ou différente. Par exemple, le caractère 'ã', a avec tilde, peut être saisi directement comme tel (U+00E3) ou écrit comme une combinaison (U+0061 U+0303). L'affichage à l'écran avec OpenOffice Writer (version 3.2.1) est équivalent, mais la fonction rechercher appliquée à l'un de ces caractères ne permet pas de trouver l'autre ; le logiciel Notepad++ (version 5.8.6) ne permet pas d'afficher correctement les versions combinées des caractères à l 'écran. La fonction rechercher ne permet pas non plus de retrouver toutes les occurrences d'un même caractère.

Caractères tifinagh

Nous complétons cet état des lieux des caractères dans Unicode par un exposé de la situation des caractères tifinagh au Niger, alphabet traditionnel des touaregs tamajaqophones.

Le tamajaq fait partie des langues berbères répartis autour du Sahara et dans le nord de l'Afrique (groupe chamito-sémitique) :

— au Maroc : tarifit au nord, tamazight au centre (Moyen Atlas), tashelhiyt au sud et au sud- ouest (Haut et Anti-Atlas)

— en Algérie : taqbaylit au nord (Grande et Petite Kabylie), zénatya au sud (Mzab et Ourgla) chaouïa à l'est (Aurès), tahaggart des touaregs sahariens du Hoggar.

— au Mali: tamajaq de l'Adrar

— au Niger : tamajaq au nord (Aïr), au centre (vallée de l'Azawagh) et à l'ouest (le long du fleuve Niger).

Il existe également de petites communautés berbères en Mauritanie, en Tunisie ou encore en Libye [Aghali-Zakara 1996].

Suite à une proposition marocco-franco-canadienne [Andries 2004] des caractères tifinagh ont été introduits au sein du répertoire Unicode [Unicode 2005], mais il apparaît qu'ils ne sont complètement adaptés à la population touarègue nigérienne utilisatrice d'alphabets tifinagh de manière traditionnelle. Au Niger, coexistent principalement deux alphabets traditionnels correspondant aux zones géographiques de l'Aïr et de l'Azawagh. Ces alphabets transcrivent 21 consonnes et la voyelle 'a' et diffèrent en ce qui concerne trois signes [Modi 2007]. De plus, ils se distinguent de l'alphabet officiel à base latinisée (voir 1.4) par l'absence de notation des consonnes emphatiques.

Valeur phonétique	Aïr	Azawagh
ɣ	ù	Q
Q	q	X
X	q	Ö

Table 2 : caractères divergents entre l'Aïr et l'Azawagh

De décembre 2001 à mars 2002, les caractères tifinagh ont été rénovés au Niger par un comité de linguistes spécialistes du tamajaq[8] [Elghamis 2003]. Cet alphabet fait la synthèse des caractères de l'Aïr et de l'Azawagh[9], de l'alphabet à base latine en usage pour la transcription (voir 1.4). Les linguistes ont effectué des choix là où il y avait des divergences entre les tifinaghs de l'Aïr et de l'Azawagh et fait des

---

[8] Ce comité était piloté :

- à Paris par Mohamed Aghali-Zakara ;

- à Agadez par Ghoubeïd Alojaly, assisté de Emoud Salekh, Ahmed Amessalamine, Ahmed Moussa Nounou,

Mohamed Adendo, Alhour Ag Analoug, Abda Annour, Aghali Mohamed Zodi, Moussa Ag Elekou ;

- à Niamey par Ramada Elghamis, avec Aghali Zennou, Ibrahim Illiasso, et Adam Amarzak. 9 Par conséquent, les signes 'J̌' et 'ǧ' en sont absents.

propositions pour la notation des voyelles ; les consonnes "v" et "p", utiles pour noter les emprunts, ont été ajoutées ; les signes notant les consonnes emphatiques 'ḍ', 'ḷ', 'ṣ', 'ṭ', 'ẓ' ont simplement été construits en ajoutant un point sous le signe tifinagh notant respectivement 'd', 'l', 's', 't', 'z'. Il apparaît que l'apprentissage traditionnel de cette écriture au sein des villages facilite l'acquisition du système officiel lors de l'entrée à l'école. Par ailleurs, il existe des publications (journaux, livres) utilisant cet alphabet.

Mais certains caractères de cet alphabet sont absents de l'alphabet tifinagh du standard Unicode

[Unicode 2005], ou bien ont des interprétations différentes.

Caractères latins	**Tifinagh API**	**Unicode**	
a	a	U+2D30	Tifinagh letter ya
a	à	U+2D30 U+0306	Tifinagh letter ya combining breve
b	b	2D40	Tuareg letter yab
c	ç	—	—
d	d	U+2D39	Tifinagh letter yadd
ḍ	D	U+2D39 U+323	Tifinagh letter yadd combining dot below
e	e	—	—
ə	é	—	—
f	f	U+2D3C	Tifinagh letter yaf
g	g	U+2D36	Tifinagh letter yaj
ɣ	ù	U+2D58	Tifinagh letter yagh
h	h	U+2D42	Tifinagh letter yah
i	i	U+2D62	Tifinagh letter yay
j	j	U+2D4C	Tifinagh letter tuareg yazh
k	k	U+2D3E	Tifinagh letter tuareg yak
l	l	U+2D4D	Tifinagh letter yal
ḷ	L	U+2D4D U+323	Tifinagh letter yal combining dot below
m	m	U+2D4E	Tifinagh letter yam
ŋ	è	U+2D50	Tifinagh letter tuareg yagn
n	n	U+2D4F	Tifinagh letter yan
o	o	—	—
p	p	—	—
q	q	U+2D57	Tifinagh letter tuareg yagh

r	r	U+2D54	Tifinagh letter tuareg yar
s	s	U+2D59	Tifinagh letter yas
ṣ	S	U+2D59 U+323	Tifinagh letter yas combining dot below
š	c	U+2D5B	Tifinagh letter yash
t	t	U+2D5C	Tifinagh letter yat
ṭ	T	U+2D5C U+323	Tifinagh letter yat combining dot below
u	u	—	—
v	v	—	—
w	w	—	—
x	x	U+2D46	Tifinagh letter tuareg yakh
y	y	U+2D49	Tifinagh letter yi
z	z	U+2D63	Tifinagh letter yaz
ẓ	Z	U+2D63 U+323	Tifinagh letter yaz combining dot below

Table 3 : caractères tifinagh APT et Unicode

Ce recensement fait donc apparaître l'absence de huit caractères dans le standard Unicode.

4. Futurs travaux

Les futurs travaux du projet DiLAF porteront dans un premier temps sur la correction des erreurs relevées dans les dictionnaires, et l'ajout d'entrées manquantes relatives aux mots désignés par les liens de synonymie, d'antonymie, etc.

La seconde étape consiste en un enrichissement des dictionnaires afin d'être en mesure de calculer toutes les formes fléchies des noms et adjectifs et toutes les conjugaisons des verbes.

Dans la mesure du possible une troisième étape de traduction des exemples et définitions vers une ou plusieurs autres langues sera définie afin de constituer des corpus plurilingues.

Conclusion

Le projet DiLAF établit une méthodologie de conversion de dictionnaires éditoriaux vers des formats XML. Il s'agit de créer et rendre disponibles de nouvelles ressources aux chercheurs en TAL, d'une part et de d'équiper les langues

africaines de ressources numériques nouvelles et indispensable à leur promotion, d'autre part.

La publication de ces ressources sur Internet permettra aux locuteurs de ces langues de disposer, souvent pour la première fois, d'informations linguistiquement fiables quant à l'orthographe, au lexique ou vocabulaire et à l'usage des mots de leur langue.

La tenue de ce premier atelier a permis de rassembler une dizaine de linguistes de trois pays ainsi que deux informaticiens. Les travaux menés ensemble ont fait émerger la richesse de la collaboration entre disciplines complémentaires et entre pays voisins. Les transferts de connaissance ont été riches, tant en ce qui concerne les outils techniques que sur des sujets de fond en linguistique. Les formations communes, les réalisations de chacun et les discussions ont fait émerger une synergie d'action entre les pays concernés.

Références

Aghali-Zakara, Mohamed. *Eléments de morpho-syntaxe touarègue*. CRB / GETIC, 1996.

Alphabet haoussa, *arrêté 212-99 de la République du Niger*, 1999. Alphabet kanouri, *arrêté 213-99 de la République du Niger*, 1999. Alphabet tamajaq, *arrêté 214-99 de la République du Niger,* 1999. Alphabet zarma, *arrêté 215-99 de la République du Niger*, 1999.

Andries, Patrick. *Proposition d'ajout de l'écriture tifinaghe*. Organisation internationale de normalisation. Jeu universel des caractères codés sur octets (JUC). ISO/IEC JTC 1/SC 2 WG 2 N2739, 2004.

Berment, Vincent. *Méthodes pour informatiser des langues et des groupes de langues peu dotées*. Ph.D. thesis, Université Joseph Fourier, 2004.

Chalvin, Antoine et Mangeot, Mathieu. *Méthodes et outils pour la lexicographie bilingue en ligne : le cas du Grand Dictionnaire Estonien-Français*. Actes d'EURALEX 2006, Turin, Italie, 6-9 septembre 2006, 6 p. 2006.

Elghamis, Ramada. *Guide de lecture et d'écriture en tifinagh vocalisées*. APT, Agadez, Niger, janvier 2003.

Eluerd, Roland. *La Lexicologie*. Paris, PUF, Que sais-je ? 2000.

Enguehard, Chantal. *Les langues d'Afrique de l'Ouest : de l'imprimante au traitement automatique des langues*, Sciences et Techniques du Langage, 6, p.29-50, 2009. (ISSN 0850-3923).

Haralambous, Yannis. *Fontes & codages*,O'Reilly France, 2004.

Mangeot, Mathieu. *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble I, 280 p., jeudi 27 septembre 2001.

Mangeot, Mathieu et Sérasset, Gilles et Lafourcade, Mathieu. *Construction collaborative de données lexicales multilingues, le projet Papillon*. Revue TAL, édition spéciale, Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ? (Electronic dictionaries: for humans, machines or both?) Ed. Michael Zock & John Carroll, Vol. 44:2/2003, pp. 151-176. 2003.

Mangeot, Mathieu et Thevenin, David. *Online Generic Editing of Heterogeneous Dictionary Entries in Papillon Project*. Proc. of COLING 2004, ISSCO, Université de Genève, Switzerland, 23-27 August 2004, vol 2/2, pp 1029-1035. 2004.

Mangeot, Mathieu et Chalvin, Antoine. *Dictionary Building with the Jibiki Platform: the GDEF case*. Proc. of LREC 2006, Genoa, Italy, 23-25 May 2006, pp 1666-1669. 2006.

Matoré, Georges. *La Méthode en lexicologie*. Paris, Didier, 1973.

Modi, Issouf. *Les caractères tifinagh dans Unicode*. Actes du colloque international "le libyco- berbère ou le tifinagh : de l'authenticité à l'usage pratique", p.241-254, ed. Haut Commissariat à l'amazighité (HCA). 21-22 mars 2007, Alger.

Mortureux, Marie-F. *La lexicologie entre langue et discours*. Paris, SEDES, 1997.

Streiter, Oliver et Scannell, Kevin P. et Stuflesser, Mathias. *Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies*

for Developers. Machine Translation, vol. 20 n°3, mars 2006.

The Unicode Standard 4.1, *Tifinagh*, range 2D30-2D7F, 2005.

Nous remercions spécialement M. Moukeïla Sanda, à l'initiative de ce projet, Mme Rabi Bozari, directrice de l'Institut National de Documentation, de Recherche et d'Animation Pédagogiques, Mme Rakiatou Rabé, M. Maï Moussa Maï et Mahamou Raji Adamou, linguistes, sans qui ce projet ne pourrait être mené à bien.

Le projet DiLAF est financé par le Fonds Francophone des Inforoutes de l'Organisation Internationale de la Francophonie.

http://www.inforoutes.francophonie.org/projets/projet.cfm?der_id=262

Un dictionnaire en tant que corpus : Traitements informatiques du dictionnaire raisonné berbère – français de Miloud Taïfi

Miloud Taïfi[1], Patrice Pognan[2]

[1]Université Sidi Mohamed ben Abdellah de Fès

[2]INALCO, Lalic (U. Paris Sorbonne & INALCO)

Résumé

L'entreprise que nous présentons ici possède un caractère particulier par rapport aux thématiques du congrès. En effet, à l'heure actuelle, la démarche devenue habituelle est la construction de vastes corpora avec les outils afférents qui permettent l'exploitation des données et leur mise en valeur pour des études scientifiques ultérieures (linguistique, traitement automatique des langues, littérature, sociologie, …) et pour des applications telles que l'élaboration de dictionnaires, de grammaires, de manuels d'apprentissage de la langue…

Nous avons eu un cheminement inverse, car dans notre cas, le dictionnaire était présent en premier, en cours de réalisation. Notre démarche consiste donc à utiliser une application pour produire des ressources qui permettront de nouvelles applications. A partir de travaux sur la structure de bases de données pour consigner dans le même appareil des langues indo-européennes, chamito-sémitiques et agglutinantes (langues turques et langues finno-ougriennes), nous transformons le dictionnaire berbère – français en base de données avec de nouvelles applications en dictionnairique, mais aussi dans le domaine de la langue (lexicologie, morphologie, aide à la construction de grammaires et de méthodes d'apprentissage).

1. Présentation du dictionnaire

Ce dictionnaire est une version corrigée, augmentée et remaniée de l'ouvrage intitulé « Dictionnaire Tamazight–Français (parlers du Maroc central) » publié en 1992. Le but visé étant de reconstituer le système de la langue berbère, cette nouvelle version est suffisamment différente de la première, dans plusieurs de ses aspects, pour en proposer un autre titre qui est désormais : « Dictionnaire raisonné berbère - français. Parlers du Maroc ». Cela veut dire que la masse lexicale consignée dans cet ouvrage n'est plus, comme dans le précédent, confinée dans le dialecte tamazight qui regroupe les parlers pratiqués dans le Maroc central, mais comporte aussi, comme il est expliqué ci-après, des formes de mot et/ou des significations nouvelles attestées dans d'autres zones géolinguistiques berbérophones du Maroc.

1.1. Correction

La correction a consisté à redresser les fautes, les erreurs et les maladresses aussi bien dans la partie berbère du dictionnaire que dans sa partie française. La saisie de l'ouvrage a permis de revoir avec plus de détail et d'acuité toutes les scories que comporte l'ancienne version. Les supports informatiques et leur manipulation ont été d'un grand secours. Il est cependant évident, malgré toute l'attention portée à la réécriture de l'ouvrage, que le lecteur peut trouver encore quelques coquilles ou quelques oublis.

1.2. Augmentation

La masse lexicale consignée dans cet ouvrage est augmentée de plus de 60% (presque 8200 racines contre environ 5000 dans le dictionnaire précédent). Elle regroupe les parlers pratiqués dans le Maroc central, mais comporte aussi des formes de mot et/ou des significations nouvelles attestées dans d'autres zones géolinguistiques berbérophones du Maroc.

Deux sources essentielles nous ont permis de procéder à l'augmentation de la masse lexicale répertoriée dans cet ouvrage.

- depuis 1992, la lexicographie berbère marocaine a connu un essor remarquable de par la réalisation de divers travaux sur le lexique, travaux académiques en majorité ; ce qui nous a permis de renouveler et d'augmenter le dictionnaire précédent (Oussikoum 1995, Azdoud 1997, Boumalk et Bounfour 2001, Serhoual 2002, Rahho 2005). Nous avons ainsi, en puisant, avec prudence et circonspection, dans ces travaux, enrichi le dictionnaire d'autres mots, d'autres

expressions et d'autres acceptions, en élargissant l'investigation à d'autres parlers du Maroc, appartenant aux dialectes dits tarifit ou tachelhit, en nous fondant essentiellement sur les critères de production et/ou de réception et en considérant la langue berbère du Maroc dans sa globalité, espérant, de par cette orientation, participer à l'effort collectif de la propagation du berbère auprès de tous ses locuteurs. On remarquera ainsi que nous avons renoncé à indiquer l'appartenance des formes à tel ou à tel parler ou dialecte, mettant en évidence, de cette façon, le système de la langue en elle-même et non pas les différentes et diverses performances de locuteurs. L'option de transcription adoptée dans ce dictionnaire, comme il est expliqué ci-après, renforce davantage une telle orientation ;

- la seconde source concerne les corpus de littérature orale dans toutes sa diversité : chants, poésie, proverbes, devinette, contes ... textes authentiques, les formes littéraires sont des garanties d'attestation. Nous avons exploité ainsi des documents publiés et plusieurs autres corpus collectés par des étudiants dans le cadre de leurs travaux académiques en thèse, (Amrani 2007, Kich 2007, Jarmouni 2009). Une source inestimable que la littérature, car elle conserve des mots, des expressions et des acceptions que l'usage quotidien du berbère n'actualise jamais ou du moins rarement ! L'exploitation de la littérature nous a permis aussi de diversifier les exemples du dictionnaire en y insérant d'autres chants, proverbes et devinettes.

1.3. Remaniement

Nous avons adopté de façon plus systématique une écriture phonologique et grammaticale. La première option consiste à reconstituer, quand cela est possible, les éléments constitutifs de la racine qui subissent, lors des réalisations phonétiques, des changements et des altérations dans les formes de mot ou à la frontière des constituants au niveau des séquences syntagmatiques. Les changements phonétiques ne sont pas nombreux, ils portent surtout sur des réalisations circonscrites dans l'aire linguistique du berbère. Quelques exemples suffiront à montrer de quoi il s'agit:

[1] le passage de k à š : akal > ašal « terre, sol » ;

[2] le passage de g à ž : igenna > iženna « ciel » ;

[3] le passage de l à ž : alim > ažim ou bien à r : alim > arim.

Ces changements n'altèrent en rien la structure des racines. Par contre la vocalisation des semi-consonnes a un effet corrosif. En effet « y » est souvent, dans certaines formes de mot, réalisé en la voyelle « i ». Il en est ainsi par exemple du verbe asy « prendre », dont la racine constitutive est bilitère SY, qui voit sa deuxième radicale actualisée en « i » dans certaines formes de conjugaison :

[4] usix au lieu de useyx « j'ai pris » ;

[5] tusim au lieu de tuseym « vous avez pris »

Ce qui réduit la forme verbale à une racine monolitère. Il en est de même pour la semi-consonne « w », réalisée en la voyelle « u » dans quelques contextes phoniques. Ainsi le verbe arw « enfanter, accoucher » est trilitère, mais dans certaines de ses formes conjuguées « w » est réalisé « u »:

[6] turud « tu as accouché » au lieu de turewd

ce qui réduit là aussi la racine bilitère à une monolitère.

La seconde option, qui constitue sans doute une innovation, consiste à rendre transparents et visibles dans l'écriture tous les éléments de la langue, lexicaux ou grammaticaux, constitutifs des énoncés. Une telle écriture permet ainsi l'identification des objets linguistiques tels qu'ils se présentent dans le système de la langue, indépendamment des performances aussi variées des locuteurs berbérophones.

1.4. Reconstitution du système de la langue

Le principe fondamental qui préside à la méthodologie appliquée dans ce dictionnaire est la reconstitution du système de la langue berbère, telle qu'elle se présente dans le domaine marocain. On est en effet très loin de l'époque où l'on affirmait que le berbère est constitué d'une « poussière » de parlers, chacun confiné dans une zone géographique limitée, réduite parfois à un mouchoir de poche. Plusieurs facteurs sociaux ont depuis contribué à l'ouverture des parlers et à leurs contacts avec les autres : le mouvement des populations au Maroc, une sédentarisation accélérée, les mass médias (la radio notamment), le tissu associatif défendant la cause berbère et plus récemment l'insertion du berbère dans le système éducatif ont réduit sensiblement l'étanchéité entre les parlers et les dialectes. Ajoutons aussi que les recherches académiques ont permis une connaissance plus approfondie de plusieurs parlers en révélant leurs particularismes. La reconstitution du système impose de ce fait la méthodologie appliquée dans la confection de ce dictionnaire.

Trois options en sont les plus saillantes :

- notation des formes de base selon leur phonie initiale, en procédant cependant à des renvois, quand cela est nécessaire, à des réalisations réelles particulières ;

- dissimilation des complexes phonétiques au niveau de toutes les séquences syntagmatiques dans lesquelles ils apparaissent ;

- traitement des variétés lexicales dans le cadre de la synonymie ou parasynonymie en considérant que toutes les formes de mot rapprochées de par leurs affinités de sens, appartiennent au lexique de la langue berbère, indépendamment de leur actualisation dans tel ou tel parler.

C'est ce qui justifie le qualificatif « raisonné » dont est affublé le titre de cet ouvrage.

1.5. Comparaisons

1.5.1. Comparaison avec le kabyle

Il nous a paru utile de maintenir la comparaison des données de ce dictionnaire avec le kabyle en nous référant exclusivement à Dallet (1982). La comparaison pourrait servir aux études dialectologiques. Elle nous montre que les différents dialectes se partagent un grand nombre de racines. Mais les mots formés d'une même racine ne recouvrent pas toujours les mêmes sens. Le rapprochement n'a été noté que dans des cas où les parlers berbères du Maroc et le kabyle présentent au moins un sens commun, un invariant de sens pour une même racine.

1.5.2. Comparaison avec l'arabe

Le berbère a admis beaucoup de vocables étrangers. On y trouve des mots latins, turcs, français, espagnols..., mais ce sont surtout les emprunts faits à l'arabe qui constituent la plus grande partie des apports étrangers. On sait par ailleurs que l'arabe et le berbère appartiennent à la même famille de langues : le chamito-sémitique. Les deux systèmes contiennent donc nécessairement un fond lexical commun. Dans l'état actuel, le berbère et l'arabe dialectal marocain sont en contact étroit. Il y a donc inévitablement emprunt de part et d'autre, et il n'est pas toujours aisé de statuer sur la provenance de certaines racines.

Il est évident cependant que des sous-systèmes lexicaux relatifs à des domaines particuliers : religieux, sociopolitique ..., sont empruntés à l'arabe. Mais l'examen

de l'ensemble lexical montre que les racines communes aux deux systèmes ne sont pas toutes exclusivement arabes. Les rapprochements notés dans ce dictionnaire n'indiquent donc pas l'origine des racines berbères mais qu'il y a, en synchronie, simple similitude entre le berbère et l'arabe.

1.5.3. Référence à d'autres langues

Nous indiquons aussi, de façon sporadique et quand cela nous semble plausible, l'origine des emprunts faits à d'autres langues, notamment au français et à l'espagnol, au latin et au turc.

1.6. Traduction en français

La traduction d'une langue en une autre n'est pas une opération facile, et ce travail n'échappe pas aux problèmes auxquels est confrontée toute étude de lexicographie différentielle. La tâche a été d'autant plus malaisée que le berbère et le français sont deux langues qui appartiennent à des familles distinctes et représentent des cultures foncièrement différentes.

Les difficultés de la traduction relèvent en effet de deux sortes de causes :

- des causes d'ordre linguistique : les mots ne sont pas équivalents et ne recouvrent pas toujours les mêmes acceptions ;

- des causes d'ordre culturel : les langues expriment différemment les réalités environnantes.

Les exemples retenus servent à illustrer chaque sens et montrent aussi comment le berbère organise les éléments linguistiques au niveau de la chaîne. La traduction littérale est donnée dans des cas où il y a ambiguïté sémantique ou pour mettre en relief un phénomène de syntaxe. Les expressions, locutions, proverbes et pièces poétiques sont d'abord, pour la plupart, traduits littéralement pour montrer la différence entre leur sens littéral et leur signification globale.

1.7. Classification par racines et organisation des articles sous une même racine

Les racines dégagées sont classées par l'ordre alphabétique du français adapté aux phonèmes particuliers du berbère. Beaucoup de racines sont homonymes, c'est-à-dire composées des mêmes consonnes. L'homonymie concerne surtout les monolitères, les bilitères et plus rarement les trilitères. L'ordre de classification des racines homonymes est le suivant : ont été notées d'abord celles qui fournissent les

outils grammaticaux : pronoms, particules, conjonctions..., ensuite les racines verbo-nominales et, en dernier lieu, les racines qui sont exclusivement nominales.

Chaque racine dégagée est indiquée en lettres capitales. Elle constitue l'entrée-vedette d'un ou de plusieurs articles. En face de la racine, à droite sur la même ligne, sont indiqués soit l'origine, quand elle est bien établie, soit les rapprochements avec d'autres langues, notamment, ce qui est plus fréquent, avec l'arabe et/ou le kabyle.

Chaque article est introduit au début de la ligne par le signe ♦ ; sont notées immédiatement après, à l'aide de la barre oblique / les variantes phonétiques ou morphologiques.

Le signe ► introduit les sens. Les exemples viennent ensuite, précédés du signe •. La traduction est séparée de l'exemple par une simple virgule. La traduction littérale ou une note explicative sont toujours mises entre parenthèses.

2. Les traitements informatiques du dictionnaire

Les traitements informatisés réalisés (dans un environnement de programmation en Python) sur le dictionnaire permettent d'obtenir un certain nombre d'indications chiffrées. Mis « à plat », ce dictionnaire représente près de 7200 racines (ce qui constitue une augmentation d'environ 60% du nombre de racines par rapport au dictionnaire de 1992), plus de 40000 enregistrements informatiques, près de 18000 articles dont 5000 concernent des verbes simples ou dérivés et un total de 2700000 caractères. C'est aussi un ensemble de 13500 exemples et locutions berbères authentiques.

2.1. du dictionnaire au corpus

Nous avons transformé les 29 fichiers Word du dictionnaire en fichiers textes bruts codés en Unicode UTF-8 que nous avons fusionnés en un corpus relativement important de 2 700 000 caractères. A partir de ce corpus, une série de programmes doit pouvoir produire une plateforme adéquate à la construction d'une base de données comprenant le même contenu que le dictionnaire d'origine, augmenté de valeurs grammaticales calculées automatiquement.

2.2. du corpus à la base de données

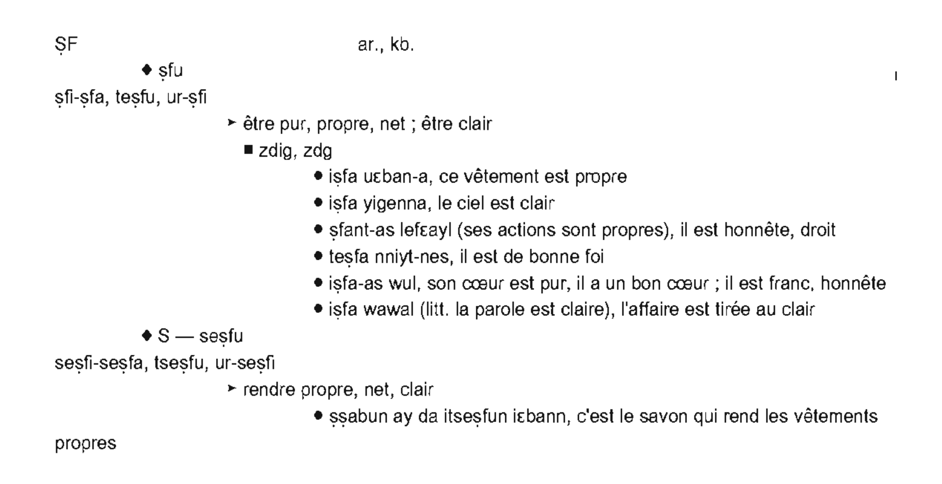
Le traitement est divisé en deux modules. Le premier a pour but de reconnaître les structures existantes du dictionnaire et le second d'enrichir ces structures de

connaissances de nature morphologique à partir des indications minimales qui sont données dans le dictionnaire.

II.2.1. Le premier module est divisé en 5 programmes (scripts en Python) qui s'enchaînent. Les structures du dictionnaire sont reconnues « à l'envers » en partant de la structure la plus profonde. Nous partons du texte en Unicode UTF-8 qui a l'apparence suivante:

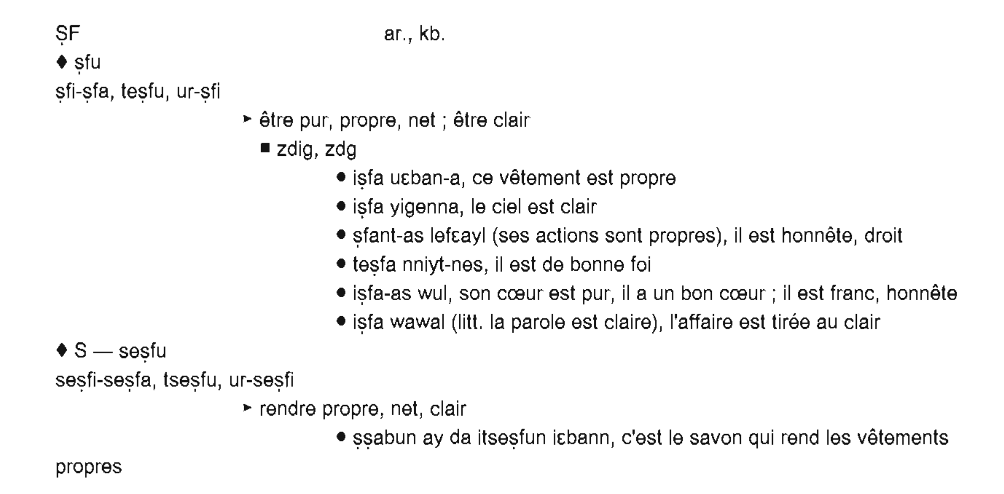
ṢF ar., kb.
♦ ṣfu
ṣfi-ṣfa, teṣfu, ur-ṣfi ► être pur, propre, net ; être clair (v. aussi : zdig, zdg) • iṣfa uɛban-a, ce vêtement est propre. • iṣfa yigenna, le ciel est clair. • ṣfant-as lefɛayl (ses actions sont propres), il est honnête, droit. • teṣfa nniyt-nes, il est de bonne foi. • iṣfa-as wul, son cœur est pur, il a un bon cœur ; il est franc, honnête. • iṣfa wawal (litt. la parole est claire), l'affaire est tirée au clair.
♦ S — seṣfu
seṣfi-seṣfa, tseṣfu, ur-seṣfi ► rendre propre, net, clair. • ṣṣabun ay da itseṣfun iɛbann, c'est le savon qui rend les vêtements propres.

Le premier programme repère les exemples grâce au signe typographique qui les introduit et les décale vers la droite pour amorcer une structuration du texte:



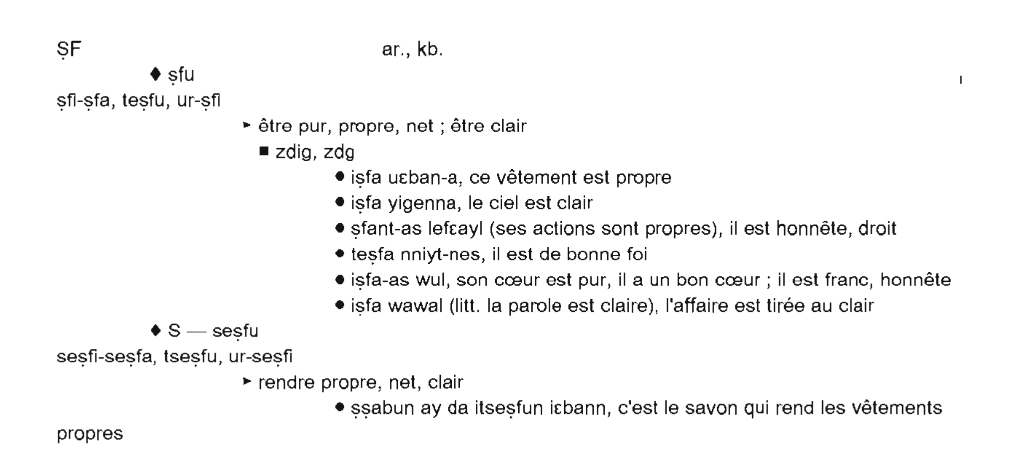
ṢF ar., kb.
♦ ṣfu
ṣfi-ṣfa, teṣfu, ur-ṣfi
► être pur, propre, net ; être clair
■ zdig, zdg
• iṣfa uɛban-a, ce vêtement est propre
• iṣfa yigenna, le ciel est clair
• ṣfant-as lefɛayl (ses actions sont propres), il est honnête, droit
• teṣfa nniyt-nes, il est de bonne foi
• iṣfa-as wul, son cœur est pur, il a un bon cœur ; il est franc, honnête
• iṣfa wawal (litt. la parole est claire), l'affaire est tirée au clair
♦ S — seṣfu
seṣfi-seṣfa, tseṣfu, ur-seṣfi
► rendre propre, net, clair
• ṣṣabun ay da itseṣfun iɛbann, c'est le savon qui rend les vêtements propres

Le second programme traite le niveau des significations et en particulier les problèmes d'absence de signification provoqués par des mentions telles que « même sens que préc. » en recherchant et en dupliquant le sens précédent. Il effectue une autre tâche importante: il extrait et ordonne les éléments de synonymie, signalés par un carré:



ṢF ar., kb.
♦ ṣfu
ṣfi-ṣfa, teṣfu, ur-ṣfi
‣ être pur, propre, net ; être clair
■ zdig, zdg
● iṣfa uɛban-a, ce vêtement est propre
● iṣfa yigenna, le ciel est clair
● ṣfant-as lefɛayl (ses actions sont propres), il est honnête, droit
● teṣfa nniyt-nes, il est de bonne foi
● iṣfa-as wul, son cœur est pur, il a un bon cœur ; il est franc, honnête
● iṣfa wawal (litt. la parole est claire), l'affaire est tirée au clair
♦ S — seṣfu
seṣfi-seṣfa, tseṣfu, ur-seṣfi
‣ rendre propre, net, clair
● ṣṣabun ay da itseṣfun iɛbann, c'est le savon qui rend les vêtements propres

Le troisième programme isole les différentes racines en les séparant les unes des autres par une ligne vide et met en évidence les articles du dictionnaire, c'est-à-dire les différents mots relevant de la racine en question:



ṢF ar., kb.
♦ ṣfu
ṣfi-ṣfa, teṣfu, ur-ṣfi
‣ être pur, propre, net ; être clair
■ zdig, zdg
● iṣfa uɛban-a, ce vêtement est propre
● iṣfa yigenna, le ciel est clair
● ṣfant-as lefɛayl (ses actions sont propres), il est honnête, droit
● teṣfa nniyt-nes, il est de bonne foi
● iṣfa-as wul, son cœur est pur, il a un bon cœur ; il est franc, honnête
● iṣfa wawal (litt. la parole est claire), l'affaire est tirée au clair
♦ S — seṣfu
seṣfi-seṣfa, tseṣfu, ur-seṣfi
‣ rendre propre, net, clair
● ṣṣabun ay da itseṣfun iɛbann, c'est le savon qui rend les vêtements propres

Le quatrième programme traite le niveau du mot où il regroupe toutes les informations morphologiques et met entre crochets les remarques à la fin de la ligne, p. ex. « [même racine que la précédente ?] »:

ṢF ar., kb.
♦ ṣfu
ṣfi-ṣfa, teṣfu, ur-ṣfi ▸ être pur, propre, net ; être clair (v. aussi : zdig, zdg)
• iṣfa uɛban-a, ce vêtement est propre
• iṣfa yigenna, le ciel est clair
• ṣfant-as lefɛayl (ses actions sont propres), il est honnête, droit
• teṣfa nniyt-nes, il est de bonne foi
• iṣfa-as wul, son cœur est pur, il a un bon cœur ; il est franc, honnête
• iṣfa wawal (litt. la parole est claire), l'affaire est tirée au clair
♦ S — seṣfu
seṣfi-seṣfa, tseṣfu, ur-seṣfi ▸ rendre propre, net, clair.
• ṣṣabun ay da itseṣfun iɛbann, c'est le savon qui rend les vêtements propres

Lorsque l'entrée lexicale est de nature nominale, le premier mot, celui qui constitue l'entrée, peut porter une indication sur l'état d'annexion: « ♦ uṣbiḥ (wu), ».

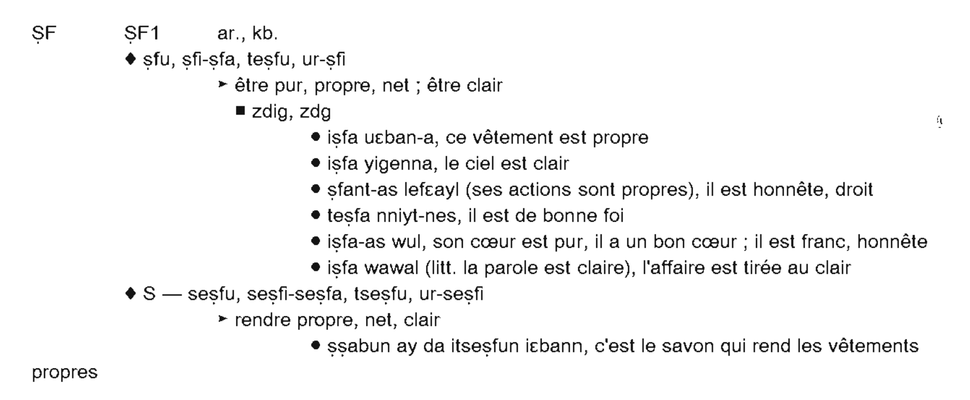
Le cinquième programme ne traite que les indications d'occurrence d'une racine dans d'autres langues, par exemple en kabyle et/ou en arabe:

ṢF ar., kb.
♦ ṣfu, ṣfi-ṣfa, teṣfu, ur-ṣfi
▸ être pur, propre, net ; être clair
■ zdig, zdg
• iṣfa uɛban-a, ce vêtement est propre
• iṣfa yigenna, le ciel est clair
• ṣfant-as lefɛayl (ses actions sont propres), il est honnête, droit
• teṣfa nniyt-nes, il est de bonne foi
• iṣfa-as wul, son cœur est pur, il a un bon cœur ; il est franc, honnête
• iṣfa wawal (litt. la parole est claire), l'affaire est tirée au clair
♦ S — seṣfu, seṣfi-seṣfa, tseṣfu, ur-seṣfi
▸ rendre propre, net, clair
• ṣṣabun ay da itseṣfun iɛbann, c'est le savon qui rend les vêtements propres

La sortie de ce programme constitue l'entrée du second module.

2.2.2. *Le second module* est déterminant pour pouvoir créer une base de données correspondant au dictionnaire. Il est composé de trois scripts qui calculent des informations supplémentaires.

Le premier programme participe à la connaissance quantitative du dictionnaire raisonné berbère du Maroc – français. Il indique la présence de 7183 racines au sein du dictionnaire. Mais il a une fonction beaucoup plus importante. En effet, les racines berbères présentent un taux d'homographie élevé. Pour assurer l'enchâssement correct de tables dans une base de données, il convient de les lier dans un rapport « un à plusieurs ». Pour ce faire, il est nécessaire de présenter une forme unique de racines pouvant être liées à un nombre quelconque de mots en découlant. C'est pourquoi nous avons dû numéroter les racines pour assurer leur unicité. Dans les bases de données correspondantes, la racine nue est présentée à l'utilisateur, mais la racine numérotée assure l'intégrité et le bon fonctionnement de la base:


ṢF ṢF1 ar., kb.
♦ ṣfu, ṣfi-ṣfa, teṣfu, ur-ṣfi
➤ être pur, propre, net ; être clair
■ zdig, zdg
• iṣfa uɛban-a, ce vêtement est propre
• iṣfa yigenna, le ciel est clair
• ṣfant-as lefɛayl (ses actions sont propres), il est honnête, droit
• teṣfa nniyt-nes, il est de bonne foi
• iṣfa-as wul, son cœur est pur, il a un bon cœur ; il est franc, honnête
• iṣfa wawal (litt. la parole est claire), l'affaire est tirée au clair
♦ S — seṣfu, seṣfi-seṣfa, tseṣfu, ur-seṣfi
➤ rendre propre, net, clair
• ṣṣabun ay da itseṣfun iɛbann, c'est le savon qui rend les vêtements propres

Le deuxième programme est le programme clé de toute l'entreprise. Sans une analyse exacte des catégories lexicales, le projet de construction automatique de la base de données correspondante serait irréalisable. Son but est double. Il doit pouvoir assurer la construction, à partir de la BD, de deux types de dictionnaires résultants, l'un classé par les racines, l'autre classé par les mots. Il doit pouvoir aussi enrichir toutes les informations morphologiques.

Le classement fondamental à introduire est la perspective des grammaires chamito-sémitiques qui classent le matériau lexical en deux classes de dérivation: verbo-nominale (racine donnant des verbes) et nominale (sans verbe dans la dérivation).

2.2.2.a. *Calcul des dérivations et de l'échelle de classement*

Le programme calcule une échelle de valeurs de classement des mots sous la racine:

- les verbes simples reçoivent un grammatème composé des valeurs: dérivation verbo-nominale (VN), verbe (V) et verbe simple (échelle de valeurs mise à: 1.1).

- les verbes composés reçoivent le grammatème: "VN V 1.2"

- les nominaux issus de verbes: "VN SUBST 1.3. "

- et les formes nominales (substantifs et adjectifs) prennent la marque de la dérivation nominale (N) avec les valeurs (SUBST ou ADJ) et (2.):

"N SUBST 2".

Tous les verbes sont accompagnés de leurs quatre formes principales, p. ex.:

verbe « faire » : "sker sker teskar ur-skir".

C'est cette échelle de classement des mots sous la racine qui permet par la suite de produire un dictionnaire par racines et/ou un dictionnaire par mots avec référence aux racines.

2.2.2.a. *Calcul des valeurs morphologiques*

Le programme détermine les catégories lexicales – verbes simples (1.1.), verbes dérivés (1.2.) et nominaux (substantifs et adjectifs) issus d'un verbe (1.3.) ou non (2.). Il calcule toutes les formes d'annexion sur la base des formes libres.

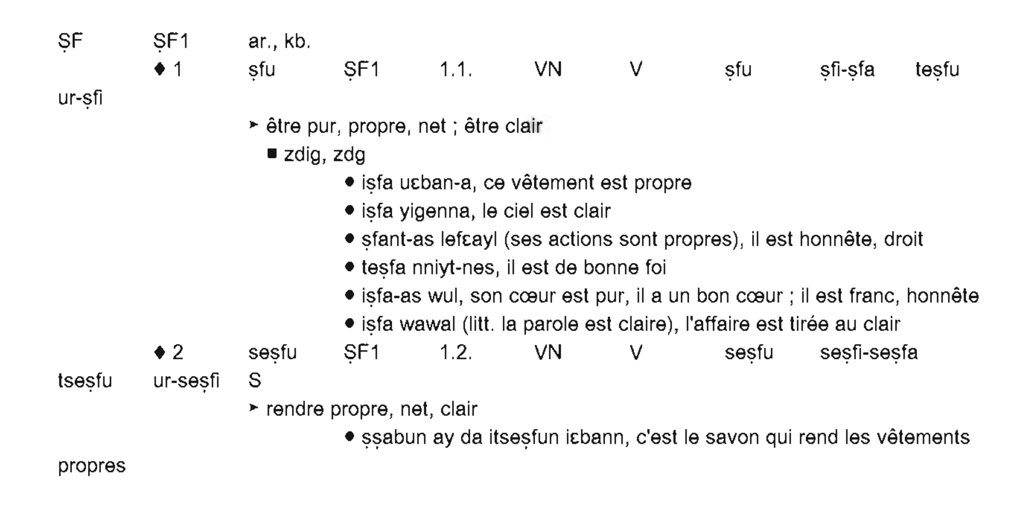
Pour pouvoir organiser au mieux ces calculs, les résultats sont consignés dans un vecteur à 24 positions représenté ci-dessous :

0	racine numérotée – losange – entrée n°: ṢBḤ2♦ṣbiḥ1	12	4ème forme verbale – prétérit négatif / accompli négatif
1	racine originelle	13	valeur du verbe dérivé: M, S, Sm, Tu,…
2	dérivation: verbo-nominale ou nominale	14	genre
3	numérotation de l'entrée au sein d'une racine: « ♦ 1 »	15	masculin singulier, état libre
4	entrée lexicale (mot)	16	masculin singulier, état annexé
5	racine numérotée	17	masculin pluriel, état libre
6	niveau du mot sous la racine: 1.1. , …, 2.	18	masculin pluriel, état annexé

7	dérivation en symboles: VN ou N	19	féminin singulier, état libre
8	catégorie lexicale: V, SUBST ou ADJ...	20	féminin singulier, état annexé
9	1ère forme verbale – aoriste	21	féminin pluriel, état libre
10	2ème forme verbale – prétérit / accompli	22	féminin pluriel, état annexé
11	3ème forme verbale – aoriste intensif / inaccompli	23	commentaires du niveau morphologique entre [] ou ()

La zone grisée marque des champs qui ne sont pas imprimés: ils servent à la construction des bases de données afférentes au dictionnaire. La zone couleur saumon est afférente aux verbes, celle en bleu clair concerne les substantifs et les adjectifs. Les zones blanches sont communes à tous les enregistrements.

Nous donnons ci-dessous l'état de l'extrait que nous avons suivi à travers cet article après classification et expression totale des valeurs morphologiques. Cet extrait ne possédant que des verbes, nous ajouterons ici un autre exemple présentant d'autres catégories lexicales.



ṢF ṢF1 ar., kb.
♦ 1 ṣfu ṢF1 1.1. VN V ṣfu ṣfi-ṣfa teṣfu ur-ṣfi
➤ être pur, propre, net ; être clair
■ zdig, zdg
● iṣfa uɛban-a, ce vêtement est propre
● iṣfa yigenna, le ciel est clair
● ṣfant-as lefɛayl (ses actions sont propres), il est honnête, droit
● teṣfa nniyt-nes, il est de bonne foi
● iṣfa-as wul, son cœur est pur, il a un bon cœur ; il est franc, honnête
● iṣfa wawal (litt. la parole est claire), l'affaire est tirée au clair
♦ 2 seṣfu ṢF1 1.2. VN V seṣfu seṣfi-seṣfa tseṣfu ur-seṣfi S
➤ rendre propre, net, clair
● ṣṣabun ay da itseṣfun iɛbann, c'est le savon qui rend les vêtements propres

Nous donnons ci-dessous un extrait comprenant un verbe et un adjectif:

ṢBḤ ṢBḤ2
♦ 1 ṣbiḥ ṢBḤ2 1.1. VN V ṣbiḥ ṣbiḥ teṣbiḥ ur-ṣbiḥ
➤ être beau, joli ; être agréable, charmant
■ fulki, flk
■ ġuda, ġd
■ iẓill, ẓl
● ur teṣbiḥ illi-s, sa fille n'est pas belle
♦ 2 uṣbiḥ ṢBḤ2 1.3. VN ADJ uṣbiḥ wuṣbiḥ uṣbiḥn wuṣbiḥn tuṣbiḥt tuṣbiḥt tuṣbiḥin tuṣbiḥin
➤ beau, joli ; agréable, charmant

Le troisième et dernier programme a pour but de traiter les exemples et de produire l'exemple berbère et sa traduction française accompagnés éventuellement d'une traduction littérale. Il est en pleine reconstruction, afin d'obtenir le maximum de solutions informatiques nous évitant les interventions manuelles sur la base de données induite, celles-ci étant toujours très longues.

Le problème vient du fait que l'exemple berbère est séparé de la traduction par une virgule qui est un symbole de ponctuation hautement ambigu (il suffit de se reporter à l'extrait que nous avons donné tout au long de l'article pour s'en persuader). Nous sommes donc en train de faire une étude sur la ponctuation au sein des exemples, d'une part et avons préparé une segmentation des exemples analysés au niveau des graphèmes, d'autre part.

Cette analyse est basée sur la reconnaissance des signes particuliers du berbère: „ č š ž ḍ ġ ḥ ṛ ṣ ṭ ẓ ɛ " et de ceux du français: „ ' à æ é è ê ï î ô œ ù û ç ". Elle est complétée par la recherche de prépositions, conjonctions, particules,… telles que „d", „n", „s", … en berbère.

Nous présentons ci-dessous les résultats de cette dernière analyse:

● ur telli ṣṣabt aseggwas-a is ur iwwit unẓar b
cette année f
il n'y a pas de bonne récolte parce qu'il n'a pas plu suffisamment. f
● ur telli ṣṣabt aseggwas-a is ur iwwit unẓar cette année, il n'y a pas de bonne récolte parce qu'il n'a pas plu suffisamment.

Ces résultats seront combinés avec ceux obtenus par l'étude de la ponctuation:

seg ṣṣbaḥ	depuis le matin
ṣṣbaḥ zikk	le matin de bonne heure
ur teṣbiḥ illi-s	sa fille n'est pas belle
iṣebben taqebbut-nes	il a lavé au savon sa djellaba
iṣebben, meskin !	il est tout pâle, livide, le pauvre !
lqalb n ṣṣabun	morceau de savon

3.3. du corpus structuré aux applications

En premier lieu, à partir de la plateforme obtenue (corpus structuré), il est possible de construire la base de données que l'on veut: individuelle (Access, SQLite,…) ou sur serveur par utilisation de XML ou d'un SGBD (MySql, PostgreSql,…).

Une autre application immédiate a été la construction de deux lexiques annexés au dictionnaire. Il s'agit d'un lexique allant du mot berbère vers la racine berbère:

iziġ: ZX1	žawr: ŽWR2	žewwa: ŽW1	žžel: ŽL2
izikr: ZKR3	žawr: ŽWR3	žewweq: ŽWQ1	žželbana: ŽLBN1
izim: ZM2	žber: ŽBR1	žeyyef: ŽYF1	žželf: ŽLF1
izimmer: ZMR4	žber: ŽBR2	žeyyer: ŽYR1	žželṭiṭa: ŽLṬ1
izimr: ZMR4	ždeb: ŽDB1	žeεžeε: Žε1	žženžar: ŽNŽR1
izinfer: ZNFR2	ždeb: ŽDB2	žeεleq: ŽεLQ1	žžent: ŽN4
izl: ZL1	žebben: ŽBN1	žeεṭeṭ: ŽεṬ1	žžerda: ŽRD2
izli: ZL8	žebber: ŽBR3	žfel: ŽFL1	žžerḍa: ŽRD2
izm: ZM2	žedder: ŽDR1	žġem: ŽĠM1	žžernan: ŽRN1
izreẓẓi: ZRẒ1	žeffef: ŽF3	žhaž: ŽHŽ1	žžert: ŽR3
izri: ZR11	žeffen: ŽFN1	žhed: ŽHD1	žžerṭ: ŽR12

et d'un autre partant d'une signification française vers une ou plusieurs racines berbères. Ce dernier lexique est important parce qu'il représente déjà un embryon de dictionnaire français – berbère et parce qu'il permet une utilisation du dictionnaire à partir du français. Nous sommes en train de terminer sa mise en forme:

abreuver (du bétail à un point d'eau): ĠB4
abreuvoir: SRŽ2, ŠRŽ3
abri: DRG1, DRY3, NTL1, SFL2, SNFY1
  ~ à bétail: NWL2
  ~ pour se protéger: ŽY1
  ~ sous roche: FR10
  se mettre à l'~: DRY3
  tenir à l'~ des regards: ḤŽB1
abricot: MŠ8
abricotier: MŠ8
abrité: SDRY1
abriter (s'~): DRY3, NTL1, SDRY1, SFL2
accablé: LF3, NBR1
  ~ de malheurs: ḤDŠ1
  ~ de soucis: ĠF1
  ~ par les fortes chaleurs de l'été: GLF3
accablement: DƐQ1, ŠHM1
accabler: DƐQ1, ḌR5, NBR1, NĠ2, QHR1, ŠHM1
  ~ de problèmes: ŠF3
accalmie: NFŽ2
accéder (aux désirs): NƐM1
accélérer: DMR1, MR2, Z3
accentuer (s'~): NHM1

D'autres applications sont rendues possibles par la plateforme ou par la base de données correspondante:

- le projet essentiel est la préparation du matériau nécessaire à la construction d'un dictionnaire français – berbère du Maroc central en renversant la masse lexicale et les informations qui y sont liées.

- l'obtention d'un ensemble d'environ 13500 exemples et locutions, dénommé «exemplier Taïfi » pour lequel nous préparons des outils de consultation :

- iferreḍ g isekwla alliy qqurn, il a négligé les arbres (fruitiers), et ils ont séché (ils sont devenus secs).
- ituferreḍ g umuḍin alliy immut, le malade a tellement manqué de soins qu'il en est mort.
- la itferfid xef teflewt g tllas, il cherche la porte à tâtons dans le noir.
- ar itferfid illi-s n ɛemmi-s ddaw n yiḥerbel, il caressait discrètement sa cousine sous la couverture.
- ifreg s uzeggwar i yišerwan, il a construit un enclos pour agneaux avec du jujubier sauvage.
- ifreg i waman, il a mis une digue pour dévier le cours d'eau.
- ifreg-as ɛemmi-s alliy yusy ixf-nes, son oncle paternel l'a pris sous sa tutelle jusqu'à sa majorité.
- ad ax ifreg ṛebbi seg wallen, que Dieu nous protège du mauvais œil (des yeux).
- freg i wussan-nek (protège tes jours), préserve ton âme du châtiment (conseil donné au calomniateur).

- un ensemble d'environ 5000 verbes qui pourra servir de base à un conjugueur automatique:

- plusieurs dictionnaires et lexiques:

dictionnaire classé par mots avec indication de la racine,

dictionnaire inverse (a tergo),

lexique terminologique avec classement thématique (plantes, animaux, artisanat, ...),...

En conclusion, la richesse d'un grand dictionnaire justifie son usage comme source de données informatisées avec une multiplicité d'applications très satisfaisante.

ṣaḥa	ṢḤ2	1.1.	VN	V	ṣaḥa	ṣaḥa	tṣaḥa	ur-ṣaḥa	
ṣleb	ṢLB1	1.1.	VN	V	ṣleb	ṣleb	teṣlab	ur iṣlib	
tuṣləb	ṢLB1	1.2.	VN	V	tuṣləb	tuṣləb	ttuṣlab	ur-tuṣlib	Tu
ṣəlleb	ṢLB2	1.1.	VN	V	ṣəlleb	ṣəlleb	tṣəllab	ur-ṣellib	
ṣleḥ	ṢLḤ1	1.1.	VN	V	ṣleḥ	ṣleḥ	teṣlaḥ	ur-ṣliḥ	
ṣleḥ	ṢLḤ2	1.1.	VN	V	ṣleḥ	ṣleḥ	teṣlaḥ	ur-ṣliḥ	
ṣalḥ	ṢLḤ2	1.1.	VN	V	ṣalḥ	ṣalḥ	tṣalaḥ	ur-ṣalḥ	
tuṣləḥ	ṢLḤ2	1.2.	VN	V	tuṣləḥ	tuṣleh	ttuṣlaḥ	ur-tuṣliḥ	Tu
mṣalaḥ	ṢLḤ2	1.2.	VN	V	mṣalaḥ	mṣalaḥ	temṣalaḥ	ur-mṣalaḥ	M
səmṣalaḥ	ṢLḤ2	1.2.	VN	V	səmṣalaḥ	səmṣalaḥ	tsəmṣalaḥ	ur-səmṣalaḥ	Sm
ṣleḥ	ṢLḤ3	1.1.	VN	V	ṣleḥ	ṣleḥ	teṣlaḥ	ur-ṣliḥ	

Au-delà de la publication de ce dictionnaire, que nous espérons prochaine, de nombreux développements informatiques nous attendent.

Références

Achab, R. (1996). La néologie lexicale berbère (1945-1995). Peeters, Louvain.

Allati, A. (2002). Diachronie tamazighte. Université de Tétouan.

Ameur, M., Bouhjar, A., Boukhris, F., Boukous, A., Boumalk, A., Elmedlaoui, M., Iazzi, E M., Souifi, H. (2004). Initiation à la langue amazighe. Publications de l'IRCAM, Rabat.

Ameur, M., Boumalk, A. (2004). Standardisation de l'amazighe. Publications de l'IRCAM, Rabat.

Ameur, M. (2007), Emprunt et créativité lexicale en berbère. Traitement en situation d'aménagement linguistique. Thèse de doctorat d'état. Faculté des Lettres et des Sciences Humaines, Dhar El Mehraz, Fès.

Amrani, F. (2007), Le lexique berbère du corps humain (Maroc central). Approche sémantique et symbolique. Thèse de doctorat d'état. Faculté des Lettres et des Sciences Humaines, Dhar El Mehraz, Fès.

Azdoud, D. (1997), Lexique commun des Ait Hadiddou du Haut Atlas, El Jadida, Thèse de doctorat d'état. Faculté des Lettres et des Sciences humaines.

Bentolila, F. (1981). Grammaire fonctionnelle d'un parler berbère. Aït Seghrouchen d'Oum Jeniba, Selaf, Paris.

Boumalk, A. (2003). Manuel de conjugaison du tachelhit, L'Harmattan, Paris.

Boumalk, A. (1996), Morphogénèse et dynamique lexicale en berbère (tachelhit du Sud-Ouest marocain), Thèse de doctorat, Paris: INALCO.

Bounfour, A. & Boumalk, A. (2001), Vocabulaire usuel du tachelhit. Tachelhit - français, Rabat: Centre Tarik Ibn Zyad.

Chaker, S. (1995). Linguistique berbère. Etudes de syntaxe et de diachronie, Peeters, Louvain.

Dallet, J.-M. (1982). Dictionnaire Kabyle – Français, Selaf, Paris.

El Mountassir, A. (2003). Dictionnaire des verbes tachelhit – français. (parler berbère du sud du Maroc), L'Harmattan, Paris.

Jarmouni, M.-H. (2009), Anthologie analytique de la poésie berbère (tamazight) du Moyen Atlas, Fès, Thèse de doctorat. Faculté des Lettres et des Sciences Humaines, Dhar El Mehraz.

Kich, A. (2007), De la poésie orale berbère (tamazight) : typologie et mutations. Thèse de doctorat d'état. Faculté des Lettres et des Sciences Humaines, Marrakech.

Mammeri, M. (1992). Précis de grammaire berbère. Editions EnAP et Inna yas, Alger et Tizi-Ouzou.

Naït-Zerrad, K. (1995). tajeṛ ṛ umt n tmaziɣt tamirant (taqbaylit). ENAG, Alger.

Naït-Zerrad, K. (1998 -). Dictionnaire des racines berbères (formes attestées). Peeters, Louvain.

Naït-Zerrad, K (2004). Linguistique berbère et applications, L'Harmattan, Paris.

Oussikoum, B. (1995), Dictionnaire tamazight-français (parler des Ayt Wirra), Beni-Mellal, Thèse de doctorat d'état. Faculté des Lettres et des Sciences Humaines.

Pognan P. (2009). Définition d'une base de données linguistiques "à tout faire" comprenant le français - comme langue pivot - et les langues slaves de l'Ouest. In Congrès international Studia Romanistica Beliana, Banská Bystrica.

Pognan P. (2009). Définition d'un prototype général de bases de données (étude des langues slaves de l'Ouest dans une visée multilingue). In Metalanguage and Encoding Scheme Design for Digital Lexicography – Innovative Solutions for Lexical Entry Design in Slavic Lexicography, MONDILEX Third Open Workshop, Académie des Sciences Slovaque, Bratislava.

Rahho, R. (2005), Dictionnaire berbère-français. Parler des Beni-Iznassen, Fès, Thèse de doctorat. Faculté des Lettres et des Sciences Humaines, Dhar El Mehraz

Sadiqi, F. (2004). "A Grammar of Amazigh", Université Sidi Mohamed ben Abdellah-PARS, Fez.

Serhoual, M. (2002), Lexicologie, lexicographie et sémantique berbères : 1) Lexicologie amazighe 2) Dictionnaire tarifit - français. Thèse de doctorat d'état. Faculté des Lettres et des Sciences Humaines, Tétouan.

Taïfi, M. (1991). Dictionnaire tamazight - français. Parlers du Maroc central, L'Harmattan-Awal, Paris.

Taïfi, M. (2006). Si les berbérophones ne s'entendent pas oralement, qu'ils s'écrivent ! Pour une écriture grammaticale du berbère. In Standardisation de l'amazighe. Publications de l'IRCAM, Rabat.

Taïfi, M., Pognan, P. (2005). Langues berbères: à la recherche du système perdu. In Colloque international « Linguistique amazighe: les nouveaux horizons ». Tétouan.

On the risk of cross-language plagiarism for less resourced languages such as Amazigh

Paolo Rosso

Natural Language Engineering Lab

ELiRF, Dept. SIC, Universidad Politécnica de Valencia, Spain

http://www.dsic.upv.es/grupos/nle/

prosso@dsic.upv.es

Abstract

The exact population of Amazigh speakers is hard to be said since most North African countries do not record language data. What is a fact is that Amazigh is a less resourced language with a very low degree of representation on the Web. In a society where information in multiple languages is available on the Web, cross-language plagiarism is occurring every day with increasing frequency, especially for less resourced languages. Potentially this could be the case of Amazigh. The lack of resources, such as Amazigh-Arabic and Amazigh-French, makes the detection of cross-language plagiarism a real challenge. This paper gives an overview of what plagiarism is and what are the available plagiarism detection tools, as well as the state-of-the-art plagiarism detection systems, focusing especially on the case where plagiarism occurs across languages. Special emphasis will be given to cross-language plagiarism in less resourced languages such as Amazigh.

1. Introduction

A relatively sparse population speaking a group of closely related and similar languages and dialects extends across the Atlas Mountains, the Sahara and the northern part of the Sahel in Morocco, Algeria, Niger, Mali, Tunisia, Libya, and the Siwa oasis area of Egypt[1]. There is a movement among speakers of the closely related languages to unite them into a single standard language: Amazigh. The exact population of Amazigh speakers is not easy to estimate, since most North African countries do not record language data. A survey included in the official Moroccan census of 2004 and published by several Moroccan newspapers[2] gave the following figures: 34% of people in rural regions spoke Amazigh and 21% in

[1] http://en.wikipedia.org/wiki/Amazigh_language

[2] http://www.bladi.net/marocain-berbere.html

urban zones did, the national average would be 28.4% or 8.52 millions. However, it is possible that the survey asked for the language "used in daily life" which would result of course in figures clearly lower than those of native speakers. Others estimate that the total number of speakers of Amazigh in the Maghreb appears to lie anywhere between 16 and 25 (30 millions if Sahel and the Siwa oasis are included) whose vast majority are concentrated in Morocco and Algeria.

In recent years, due to the large amount of text available on the WWW, plagiarism cases have increased. Moreover, in a society where information is available on the Web in multiple languages, cross-language plagiarism cases are also common, especially when the target language is a less resourced one (e.g. Amazigh) and the user is more likely to find the information s/he looks for in a more resourced language (e.g. English, French or Arabic). The rest of the paper is structured as follows. Section 2 defines what plagiarism is and what the different kinds of plagiarism are. The available plagiarism detection tools and the best state-of-the-art plagiarism detection systems participating at the first of plagiarism detection are also described. Section 3 is devoted to cross-language plagiarism and the first attempts to approach it. Special emphasis is given to the case where the target language is a less resourced one, such us Amazigh. Finally, in the last section some conclusions are drawn.

2. Plagiarism

Although often no distinction is made between text reuse and plagiarism and just the generic text reuse is employed, there is a narrow difference between the two. With text reuse we mean the activity whereby pre-existing written texts are used again to create a new text or version (Clough and Gaizauskas, 2009) but this does not mean that an infringement is intended: collaborative authoring (e.g. Wikipedia), news from press for newspapers (e.g. Reuters, Press Association, etc.), etc. In case the reuse of someone else's prior ideas, processes, results, or words occurs without explicitly acknowledging the original author and source then we can talk about plagiarism (IEEE, 2008). It has to be said that often plagiarism could occur, for instance, in books from narrative and events that could resemble each other to plagiarism of ideas (that is not based on words dependency) and plagiarism of ideas is nowadays (practically) impossible to be detected automatically.

Surveys of the research done in automatic plagiarism detection can be read in (Clough, 2003) and (Maurer et al., 2006). Plagiarism detection can be divided into external plagiarism detection - when, given a suspicious fragment of a document, a set of potential source documents is available - and intrinsic plagiarism detection - when the lack of a set of potential source documents makes the detection of a suspicious fragment more difficult because based only on style changes.

2.1. Plagiarism Detection Tools

Many are the tools, some of them freely available, for plagiarism detection. All of them are external plagiarism detection tools, that is, their aim is to find the potential source fragment plagiarism has been committed from. Of course, this is possible only if the set of potential source documents is available. Moreover, they perform well only when a simple duplicate (copy-paste) or near-duplicate (use of synonyms) plagiarism of fragment occurs. Their performance decreases dramatically in case of paraphrasing (Barrón-Cedeño et al., 2010a) or translated plagiarism across languages (Potthast et al., 2011). Therefore, if from one hand due to the large amount of information available on the Web plagiarism has increased in recent years and this makes manual plagiarism detection infeasible (Weber, 2007; Kulathuramaiyer and Maurer, 2007), from the other hand texts can be easily found, manipulated – making usage of paraphrasing or translated plagiarism - and combined. Therefore, it is important to stress that automatic plagiarism detection has only to assists experts providing them linguistic evidence for the final decision.

Below the list of ten among the most well-known plagiarism detection tools (Vallés, 2010):

i. **Turnitin**[3] is not a free plagiarism detector tool. It has been developed by John Barrie (University of Berkeley) and it is used by more than 50 universities in the world[4].

ii. **WCopyFind**[5] is a tool which was developed in 2004 by Lou Bloomfield, University of Virginia. Plagiarism is detected on the basis of the comparison of word n-grams (sequence of n words). The size of n is decided by the users although for WCopyFind (Dreher, 2007) suggest using hexagrams.

iii. **Ferret**[6] is a tool to detect plagiarism that was developed in the University of Hertfordshire (Lyon at al., 2006). It is able to analyse documents in different formats (PDF, Word and RDF). It extracts trigrams obtaining a similarity measure on the basis of the common trigrams between two documents (Malcom and Lane, 2008).

---

[3] http://www.turnitin.com/

[4] Digital solutions for a new era in information. 2004. iparadigm: http://www.iparadigms.com

[5] http://plagiarism.phys.virginia.edu/

[6] http://homepages.feis.herts.ac.uk/_pdgroup/

iv. **CopyCatch**[7] is a tool designed by CFL Software. It is possible to calculate the similarity between two complete documents or some of its sentences. CopyCatch needs to have as input the document in order to investigate if some of its parts have been plagiarised. It succeeds in detecting the similarity also in case of simple paraphrasing: insertions, deletions or change in the order of the words. It works in different languages.

v. **iThenticate**[8] is a plagiarism detection service for preventing from Web-based plagiarism, content verification and intellectual property copyright. Given a document, it compares it against its large data base. A report is provided to the user in case a similarity is found with other(s) document(s).

vi. **Plagiarism Checker**[9] is a Web application which has been developed by the Department of Education of the University of Maryland. Its aim is to detect whether a text is suspicious to be copied. The suspicious text needs to be introduced and the application checks for similar texts using the API of Google. It is free and fast but, as most of these tools, it is quite unlikely to find the source text in case of paraphrasing or translated plagiarism.

vii. **Pl@giarism**[10] is a freely available tool that has been developed by the Law Faculty of the University of Maastricht in order to detect plagiarism cases in the essays of their students. Pl@giarism is a simple application for Windows which determines the similarity between two documents on the basis of the comparison of their trigrams. It returns a table with similarity percentages between the suspicious document and its similar documents.

viii. **DOC Cop**[11] is a freely available tool. It returns acceptable results especially if the comparison of the suspicious document is made against a smaller data base than the Web (Scaife, 2007). A report is sent by email and those fragments suspicious to be plagiarised are highlighted.

ix. **EVE2**[12] (Essay Verification Engine) is a tool developed by Canexus. EVE2 allows checking if students have plagiarised parts of their essay from the Web. It returns the links to the Web pages plagiarism is likely to have been committed from. Unfortunately it seems to be quite slow: Dreher (Dreher, 2007) carried out an

[7] http://csoftware.com/
[8] http://www.ithenticate.com/
[9] http://www.dustball.com/cs/plagiarism.checker/
[10] http://www.plagiarism.tk/
[11] http://www.doccop.com/
[12] http://www.canexus.com/

experiment in order to detect possible plagiarised texts in just 16 pages, containing 7,300 words, of a M.Sc. thesis and the tool took 20 minutes to process them.

x. **MyDropBox**[13] is an online service whose aim is to help the detection of plagiarism. The reports that the tool returns are quite well structured in order to highlight the links with the sources of the Web where plagiarism is likely to have been committed (Scaife, 2007).

2.2. External and Intrinsic Plagiarism Detection

As said previously, methods for automatic plagiarism detection can be divided in two main approaches: external plagiarism detection and intrinsic plagiarism detection.

External plagiarism detection can be considered as a task related to information retrieval. In fact, given a suspicious document *d* and a collection of potential source documents *D*, the task is to detect the plagiarised sections in *d* (if there are any), and their respective source sections in *D* (Potthast et al., 2009). Up to now, researchers have paid more attention to this approach (see, for instance, the previous section on plagiarism detection tools) because obtaining the source of a potential case of plagiarism provides better linguistic evidence to help the expets (e.g. forensic linguistics) to make their final decision on whether a fragment of text has been plagiarised or not. The problem is that it is not an easy task to find the potential source of plagiarism in case the set *D* of potential source documents is the Web itself. In fact, text plagiarism is observed at an unprecedented scale with the advent of the World Wide Web (the new term of cyber-plagiarism (Comas and Sureda, 2008) has been recently introduced to refer to the copy-paste syndrome) and this is the real scenario plagiarism detection systems should consider. In terms of number of comparisons, the size of the reference data set (e.g. the Web) could be a problem from a computational point of view. Therefore, it is important to reduce the number of exhaustive comparisons only to those between fragments that are more similar. In order to solve the problem of the size of the reference data set, in (Barrón-Cedeño and Rosso, 2009) the authors described a method based on the Kullback-Leibler distance (Kullback and Leibler, 1951) for reducing the search space (the Kullback-Leibler symmetric distance measures how close the probability distributions of the reference and suspicious documents are).

Most of state-of-the-art plagiarism detection systems base their approach on the comparison of word n-grams of the fragments of the suspicious document *d* and those of the documents of the reference data set *D* (Kasprzak et al., 2009) also taking into account vocabulary expansion, for instance with Wordnet[14] (Kang et al.,

[13] http://www.mydropbox.com/
[14] http://wordnet.princeton.edu/

2006). The comparison could also be made on the basis of character n-grams (Grozea et al., 2009) where character n-grams of the suspicious documents are matched against the character n-grams of the source document (see Figure 1). A dot means that the character n-gram exists in both documents. A diagonal provides linguistic evidence of a possible plagiarism case (e.g. left corner of the graph). A diagonal together with a cluster of dots gives less evidence but a certain similarity between the two fragments of the suspicious document and the source still occurs and deserves to be manually further investigated by the forensic linguistic expert who has to make the global decision whether it is a plagiarism case or not. A similar plot approach was also employed by (Basile et al., 2009) but instead of plotting character n-grams, after a pre-process in which each word was substituted by its length (e.g. length = 6), n-grams of numbers were plotted (e.g. substituted by its length = 11 2 3 6). Once more, a dot means that the number n-gram exists in both documents, and a diagonal provides linguistic evidence of a possible plagiarism case (i.e., a sequence of words of the same length is found both in the suspicious document and in the source one).

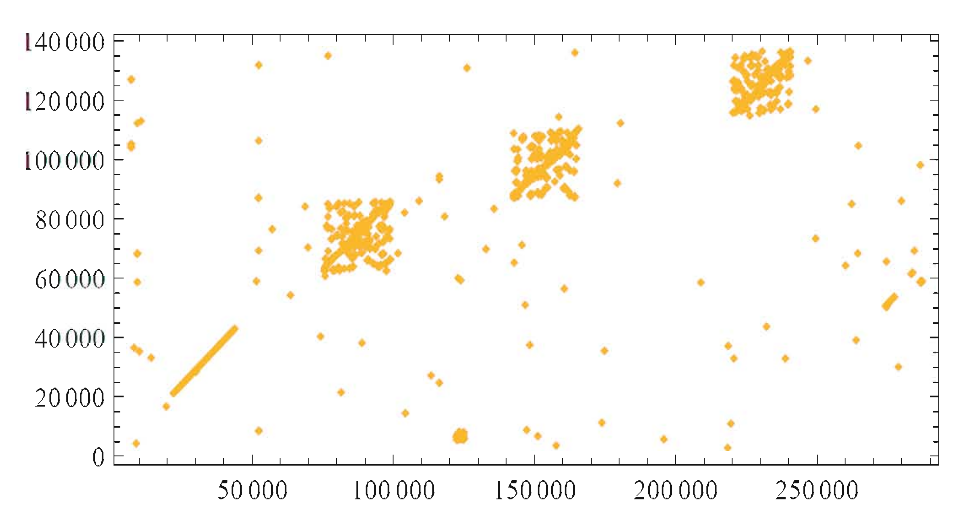


Figure 1. ENCOPLOT: visual approach for external plagiarism detection (Grozea et al., 2009)

In case of lack of the reference set of potential source documents *D*, the detection of plagiarised fragments has to rely only on changes in the writing style in the document. A person could be often able to manually identify potential cases of plagiarism by detecting text inconsistencies (unexpected irregularities through a document such as changes of style, vocabulary, or complexity are triggers of suspicion) or by resembling previously consulted material. Nevertheless, the large amount of potential source texts available nowadays makes infeasible this manual

plagiarism detection based on writing style change. In order to assist experts, automatic intrinsic plagiarism detection methods have been developed aiming to detect whether the document *d* contains text fragments written by a different author.

The features considered by these models are, among others, word length average, sentences length average, stop-words average, as well as readability and vocabulary richness (Meyer zu Eißen and Stein, 2006). The readability of a text could be measured, for instance, on the basis of the complex words used (complex words are those with three or more syllables) employing indexes such as Gunning fog or Flesch (DuBay, 2008). Figure 2 shows how linguistic evidence for plagiarism could be provided on the basis of the above measures for intrinsic plagiarism detection. In the example, two text fragments (last two columns) are compared with the all document (column named as Global). Linguistic evidence is provided with respect to the use of more complex words in the first text fragment (complexity measure of 17 vs. approx. 14). Once more, the automatic approach has the aim to simply assist the forensic linguistic expert who has to be the one making the decision. Finally, like for the external plagiarism detection, there are methods that apply character n-gram profiles to characterise an author's style and search for irregularities in the document *d* (Stamatatos, 2009).

Measure	Global	■	■
tokens	135	63	72
types	78	44	46
W. avg. freq. class			
avg. sentence length	19.28	21.00	18.0
avg. word length	4.93	5.38	4.54
Complexity measures	16.72	17.07	13.82

Figure 2. Measures for intrinsic plagiarism detection

2.3. Plagiarism Detection Competition

The development of plagiarism detection models is not new although the large amount of information available on the Web plagiarism has increased in recent years. One of the first approaches we have track of goes back to the 1970s (Ottenstein, 1976). However, after more than 30 years, no standard evaluation framework (i.e., standard text collections with documented cases of plagiarism and

evaluation measures) existed in order to compare the performance of the different plagiarism detection methods. In fact, researchers often used small and private (80% of cases (Potthast et al, 2010a)) collections of documents that cannot be freely provided to other researchers for ethical reasons. Moreover, they estimated the quality of the models by considering different evaluation measures. Therefore, with the aim of providing a standard evaluation framework on automatic plagiarism detection, together with the Webis research group of Weimar University[15] and the universities of the Aegean[16] and of Bar-Ilan[17], the first International Competition on Plagiarism detection[18] was organised. In 2011 its third edition, sponsored by Yahoo! Research Barcelona[19], will be organised again as one of the benchmarking activities of CLEF evaluation campaign[20].

In the first edition (Stein et al., 2009) two tasks were organised: external plagiarism detection and intrinsic plagiarism detection. The best approach for external plagiarism detection was the ENCOPLOT of (Grozea et al., 2009) and for intrinsic plagiarism detection the one of (Stamatatos, 2009). Both approaches were based on the comparison of character n-grams. The teams who participated with two of the software tools previously described (WCopyFind and Ferret) did not obtain a good performance (Potthast et al., 2009).

In the second edition no distinction between external and intrinsic plagiarism detection was made. The best approach was the one of (Kasprzak and Brandejs, 2010) that was based on word n-grams. In the first edition (Potthast et al., 2009), 10 teams participated in the external plagiarism detection task and only 4 teams in the intrinsic plagiarism detection one. In the second edition (Potthast et al., 2010a), although no distinction was made and only one plagiarism detection task was organised, many of the 18 teams that participated had their overall performance penalised because they did not solve properly (or they did not solve at all) the intrinsic plagiarism cases (30% of total plagiarism cases (Potthast et al., 2010b)). The above shows that less attention has been paid from the research community to the intrinsic plagiarism detection both because more difficult also in terms of giving linguistic evidence without a source document where the plagiarism has been committed from.

---

[15] http://www.uni-weimar.de/cms/medien/webis/home.html

[16] http://www.icsd.aegean.gr/lecturers/stamatatos/

[17] http://u.cs.biu.ac.il/~koppel/

[18] http://pan.webis.de/

[19] http://labs.yahoo.com/Yahoo_Labs_Barcelona

[20] http://clef2011.org/index.php?page=pages/labs.html

The results of the competition, as well as the description of the evaluation measured and the data set (8.4 Giga Bytes, 162,000 plagiarism cases, between training and test samples) are available at: http://pan.webis.de .

3. Cross-language Plagiarism

In a society where information is available on the Web in multiple languages, cross-language plagiarism occurs every day with increasing frequency. This behaviour was simulated in the data set of the competition where 14% of plagiarism cases were translated plagiarisms from Spanish or German into English (Potthast et al., 2010b).

3.1 Cross-language Plagiarism Detection

Cross-language plagiarism detection deals with the automatic identification and extraction of plagiarism in a multilingual setting. In this setting, a suspicious document is given, and the task is to retrieve the source documents of the suspicious fragments from a large, multilingual document collection.Up to the present time, cross-language plagiarism detection has not been approached sufficiently due to its intrinsic complexity. Whereas some commercial tools are able to perform plagiarism analyses on different languages, detecting cases of translated plagiarism is still in its infancy. In the first edition of the competition no team tried to detect the cross-language plagiarism cases (Potthast et al., 2009). In the second edition, some teams approached the problem on a monolingual basis translating the source documents in Spanish or German into English (Potthast et al., 2010a). No matter the large size of the data set (8.4 GB, 162,000 plagiarism cases) this is still a close scenario but in the open (and more realistic) scenario of the Web, it would be not feasible from a computational time point of view translating all the documents into the target language plagiarism needs to be investigated (e.g. Amazigh).

Few are the cross-language plagiarism detection approaches that have been investigated so far. Probably the two methods with a certain impact are CL-ASA (cross-language alignment-based similarity analysis) and CL-ESA (cross-language explicit semantic analysis). CL-ASA (Barrón-Cedeño et al., 2008; Pinto et al., 2009) is based on the IBM-M1 statistical machine translation model (Brown et al. 1993) and needs a parallel data set to be trained[21]. It estimates the likelihood of two text fragments of being valid translations of each other. CL-ESA is another interesting method for cross-language plagiarism detection (Potthast et al., 2008). CL-ESA intends to estimate, at semantic level, how similar two texts written in

[21] The JRC-Acquis data set was used : http://wt.jrc.it/lt/Acquis/

different languages are. This estimation is carried out on the basis of a comparable data set, such as Wikipedia (Figure 3). The CL-ASA and CL-ESA models have been compared in (Potthast et al., 2011) with the cross-language character n-gram model (CL-CNG). Despite its simplicity, CL-CNG results to be a good choice to compare text fragments across languages if they are syntactically related.

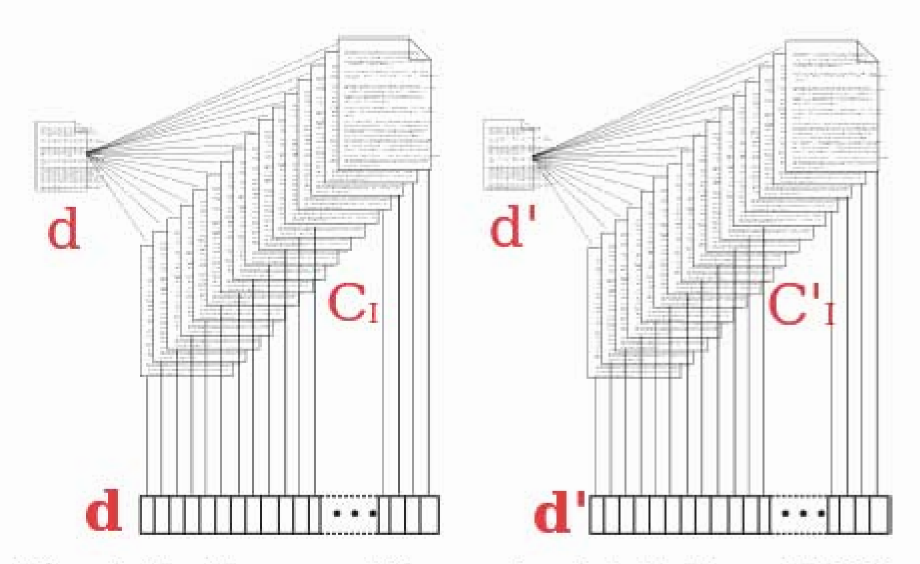


Figure 3. Cross-language explicit semantic analysis (Potthast et al., 2008).

Similarity between documents *d* and *d'* is computed on the basis of the vector space model with indexes the subset of Wikipedia common articles in both languages

3.2 Cross-language Plagiarism Detection in Less Resourced Languages

A less resourced language is that with a low degree of representation on the Web (Alegria et al., 2009). This makes not always possible to employ previous approaches such as CL-ASA and CL-ESA. CL-CNG results to be a good choice but only if the two languages are syntactically related.

If few attempts have been made to solve the problem of cross-language plagiarism detection, even less work has been done to tackle this problem for less resourced languages. One of the few works is the one of (Barrón-Cedeño et al., 2010b) on plagiarism detection across distant language pairs where the authors investigated the case of Basque, a language where, due to the lack of resources, cross-language plagiarism is often committed from texts in Spanish and English. Basque has no known relatives in the language family; however it shares some of its vocabulary

with Spanish. Therefore, the CL-CNG method based on character n-grams was investigated. CL-CNG was compared with CL-ASA and a method that approached the problem from a monolingual perspective calculating the similarity after employing a machine translation pre-process (Figure 4). The translation and monolingual similarity analysis (T+MA) performed better than the other models. As previously said, approaching the problem of cross-language plagiarism detection from a monolingual point of view after translating all the documents into the target language, would not be computationally possible in a realistic scenario such as the Web.

3.3 The Difficulty of Detecting Cross-language Plagiarism in Amazigh

Maghreb states such Morocco and Algeria have created institutions such as the Institute Royal de la Culture Amazighe (IRCAM[22]) and the Haut Commissariat à l'Amazighité (HCA[23]) in order to promote the Amazigh language. In Morocco, Amazigh has been introduced in mass media (an Amazigh television channel was launched in 2010) and in the educational system (Amazigh is taught in various Moroccan primary schools). Moreover, IRCAM during just 8 years since its creation has published more than 150 books related to the Amazigh language and culture, a number which exceeds the whole amount of Amazigh publications in the 20th century. No matter these efforts, from a computational linguistic point of view Amazigh is still a less resourced language. In fact few are the annotated large data set (e.g. (Outahajala et al., 2011)).

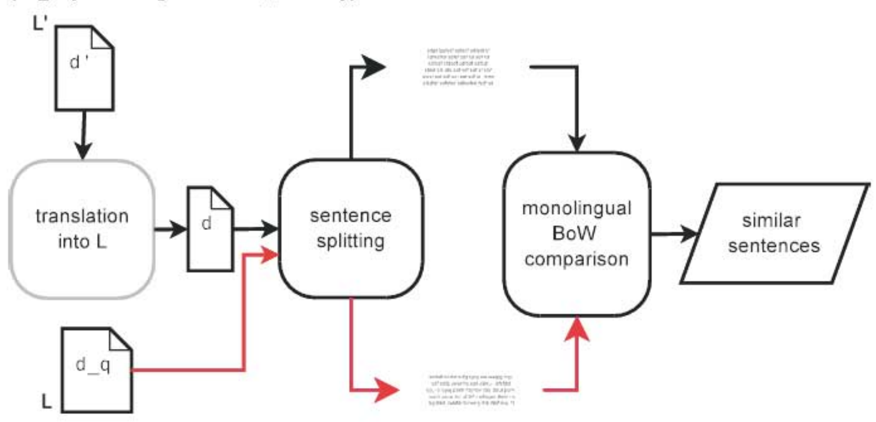


Figure 4. The translation and monolingual similarity analysis model (Barrón-Cedeño et al., 2010b)

[22] http://www.ircam.ma/
[23] http://hcamazighite.org/

The low degree of representation of Amazigh on the Web potentially could be the cause of translated plagiarism from languages such as English, Arabic or French (Figure 5) where the information could be found more easily. Amazigh is not syntactically related to English, French or Arabic and this makes not feasible using the CL-CNG model to detect cases of cross-language plagiarism cases. The lack of large parallel (in Amazigh and Arabic, French or English) and comparable data sets (e.g. Wikipedia) makes a real challenge the use of the CL-ASA and the CL-ESA models previously described. Up to the present time, IRCAM developed three parallel lexicons containing words in Amazigh and their equivalent in French[24] (Ameur et al., 2006), in French and Arabic about media[25] (Ameur et al., 2009), and in French-Arabic-English about Amazigh grammar[26] (Boumalk and Naït-Zerrad, 2009). However they are small and not parallel data sets of equivalent sentences.

Last, with respect to the possibility of employing the translation and monolingual similarity analysis (T+MA) model an automatic machine translator (French-Arabic-English into Amazigh) is needed. The possibility of having to deal with data sets in Amazigh written in both Latin or Tifinaghe characters (Figure 5) is also a further problem, although it seems that recently texts written in Tifinaghe Unicode are increasingly used.

iga usinag agldan n tussna tamaziGt yat tmrslt takadimit nna tsskr tgldit tamGrabit ittindazn i uHTTu d tnkra n tussna tamaziGt. yusa d usinag g RbaT,tamaZunt n lmGrib, ittusnta ass n 17/10/2001 s unbaD n ugllid muHmmad 6 s uDahiR agldan yusin azwil 01-01-299. iga usinag tamrslt Gur tlla tsimant taZRfant d tsmGurt.	ⵉⴳⴰ ⵓⵙⵉⵏⴰⴳ ⴰⴳⵍⴷⴰⵏ ⵏ ⵜⵓⵙⵙⵏⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ ⵢⴰⵜ ⵜⵎⵔⵙⵍⵜ ⵜⴰⴽⴰⴷⵉⵎⵉⵜ ⵏⵏⴰ ⵜⵙⵙⴽⵔ ⵜⴳⵍⴷⵉⵜ ⵜⴰⵎⵖⵔⴰⴱⵉⵜ ⵉⵜⵜⵉⵏⴷⴰⵣⵏ ⵉ ⵓⵃⵟⵟⵓ ⴷ ⵜⵏⴽⵔⴰ ⵏ ⵜⵓⵙⵙⵏⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ. ⵢⵓⵙⴰ ⴷ ⵓⵙⵉⵏⴰⴳ ⴳ ⵕⴱⴰⵟ,ⵜⴰⵎⴰⵥⵓⵏⵜ ⵏ ⵍⵎⵖⵔⵉⴱ, ⵉⵜⵜⵓⵙⵏⵜⴰ ⴰⵙⵙ ⵏ 17/10/2001 ⵙ ⵓⵏⴱⴰⴹ ⵏ ⵓⴳⵍⵍⵉⴷ ⵎⵓⵃⵎⵎⴰⴷ 6 ⵙ ⵓⴹⴰⵀⵉⵕ ⴰⴳⵍⴷⴰⵏ ⵢⵓⵙⵉⵏ ⴰⵣⵡⵉⵍ 01-01-299. ⵉⴳⴰ ⵓⵙⵉⵏⴰⴳ ⵜⴰⵎⵔⵙⵍⵜ ⵖⵓⵔ ⵜⵍⵍⴰ ⵜⵙⵉⵎⴰⵏⵜ ⵜⴰⵥⵕⴼⴰⵏⵜ ⴷ ⵜⵙⵎⵖⵓⵔⵜ.

Figure 5. French-Amazigh cross-language plagiarism: Latin (left) and Tifinaghe scripts (right).

[24] http://www.ircam.ma/fr/index.php?soc=publi&pg=5&rd=64
[25] http://www.ircam.ma/fr/index.php?soc=publi&pg=2&rd=109
[26] http://www.ircam.ma/fr/index.php?soc=publi&pg=2&rd=118

Source: http://fr.wikipedia.org/wiki/Institut_royal_de_la_culture_amazighe

4. Conclusions

Although the problem of plagiarism is well-known, not always people know what the available tools for its detection and their limitations are. Moreover, in case of less resourced languages such as Amazigh, plagiarism from other languages is more likely to occur. Automatic cross-language plagiarism detection is still in its infancy. Therefore, the detection of translated plagiarism is not possible using the available tools. This paper gives an overview of plagiarism detection and, in particular, cross-language plagiarism detection: a problem that will have to be addressed with special emphasis in the future because every time occurring more often especially for less resourced languages.

Acknowledgements

Most of what was described in this paper is the result of the joint work done together with Alberto Barrón-Cedeño in the framework of his Ph.D. and partially also with Enrique Vallés in his M.Sc. This research work was funded by the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). Last but not least, I have to thank also my Ph.D. students Mohamed Outahajala and Lahsen Abouenour for helping me with Amazigh/e .:-) Tanmmirt bzzaf!

References

Alegria, Iñaki, Mikel L. Forcada, and Kepa Sarasola, editors. (2009). Proc. of the SEPLN 2009 Workshop on Information Retrieval and Information Extraction for Less Resourced Languages, Donostia, Basque Country. University of the Basque Country.

Ameur M., Bouhjar A., Boumalk A., Elazrak N., Abdellaoui R. (2009). Vocabulaire des médias Français-Amazighe-Anglais-Arabe. Publications de l'IRCAM.

Ameur M., Bouhjar A., Elmedlaoui M., Iazzi E. (2006). Vocabulaire de la langue Amazighe. Publications de l'IRCAM.

Barrón-Cedeño A., Rosso P., Pinto D., Juan A. (2008). On Cross-lingual Plagiarism Analysis using a statistical model. In: Proc. 2nd Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, PAN-2008, Patras, Greece, July 21-24.

Barrón-Cedeño A., Rosso P. (2009) On the relevance of search space reduction in automatic plagiarism detection. In: Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 43, pp. 141-149

Barrón-Cedeño A., Vila M., Rosso P. (2010a) Detección automática de plagio: de la copia exacta a la paráfrasis. In: Panorama actual de la lingüística forense en el ámbito legal y policial: teoría y práctica. Jornadas (In)formativas de Lingüística Forense, Madrid, 21-22 October. Ed.: Euphonia Ediciones SL.

Barrón-Cedeño A., Rosso P., Agirre E., Labaka G. (2010b) Plagiarism detection across distant language pairs. In: Proc. of the 23rd International Conference on Computational Linguistics, COLING-2010, Beijing, China, August 23-27

Basile C., Benedetto D., Caglioti E., Cristadoro G., Degli Esposti M.. (2009). A plagiarism detection procedure in three steps: selection, matches and "squares". In Stein et al. (2009), pages 1–9. http://ceur-ws.org/Vol-502, pp. 19-23.

Boumalk A., Naït-Zerrad, K. (2009). Amawal n tjrrumt -Vocabulaire grammatical. Publications de l'IRCAM.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), 263–311.

Clough P. (2003). Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service. http://ir.shef.ac.uk/cloughie/papers/pasplagiarism.pdf

Clough P., Gaizauskas R. (2009). Corpora and Text Re-Use. In Lüdeling, A., Kytö, M., and McEnery, T., editors, Handbook of Corpus Linguistics, Handbooks of Linguistics and Communication Science, pages 1249—1271. Mouton de Gruyter.

Comas R., Sureda J., editors (2008). Academic cyberplagiarism, volume 10 of Digithum. Universitat Oberta de Catalunya.

Dreher H. (2007). Automatic conceptual analysis for plagiarism detection. Journal of Issues in Informing Science and Information Technology 4, pages 601-614.

DuBay W.H. (2004). The principles of readability.Impact Information,

http://www.impact-information.com/impactinfo/readability02.pdf

Grozea C., Gehl C., Popescu M. (2009). ENCOPLOT: Pairwise sequences matching in linear applied to plagiarism detection. In Stein et al. (2009), pages 1–9. http://ceur-ws.org/Vol-502, pp. 10-18

IEEE. (2008) A plagiarismFAQ. http://www.ieee.org/web/publications/rights/plagiarism_FAQ.htm, 2008. [Online; accessed 3-March-2010].

Kang N., Gelbukh A., Han S. (2006). PPChecker: Plagiarism pattern checker in document copy detection.In Proc. of the Text, Speech and Dialogue, 10th Int. Conf. TSD-2006. LNAI(4188), pp. 661–667, Springer-Verlag.

Kasprzak J., Brandejs M., Křipač M. (2009). Finding plagiarism by evaluating document similarities. In Stein et al. (2009), pages 1–9. http://ceur-ws.org/Vol-502, pp. 24-28

KasprzakJ., Brandejs M. (2010). Improving the reliability of the plagiarism detection system - Lab report for PAN at CLEF 2010. . In: Braschler M., Harman D., and Pianta E.(Eds.), Notebook Papers of CLEF 2010 LABs and Workshops, CLEF-2010, Padua, Italy, September 22-23

Kulathuramaiyer N., Maurer H. (2007). Coping With the Copy-Paste-Syndrome. In E-Learn 2007, pages 1072—1079, Quebec, CA.

Kullback S., Leibler R. (1951). On Information and sufficiency. Annals of Mathematical Statistics, 22(1):79–86.

Lyon C., Barrett R., Malcolm J. (2006). Plagiarism is easy, but also easy to detect. Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification, 1.

Malcolm J., Lane P.C.R. (2008). Efficient search for plagiarism on the web. Proc. of the Int. Conf. on Technology, Communication and Education, pp. 206-211.

Maurer H., Kappe F., Zaka B. (2006). Plagiarism - A Survey. Journal of Universal Computer Science, 12(8):1050–1084

Meyer zu Eißen S., Stein B. (2006). Intrinsic Plagiarism Detection. In Advances in Information Retrieval, Proc. of the 28th European Conf. on IR Research, ECIR 2006, LNCS(3936):565–569, Springer-Verlag.

Ottenstein K.J. (1976). An Algorithmic approach to the detection and prevention of plagiarism. ACM SIGCSE Bulletin, 8(4):30–41.

Outahajala M., Zenkouar L., Rosso P. (2011) Building an annotated corpus for Amazighe. 4ème atelier international sur l'amazighe et les TIC. Rabat.

Pinto D., Civera J., Barrón-Cedeño A., Juan A., Rosso P. (2009). A statistical approach to crosslingual natural language tasks. In: Journal of Algorithms, vol. 64, num. 1, pp. 51-60. DOI: 10.1016/j.jalgor.2009.02.005

Potthast M., Stein B., Eiselt A., Barrón-Cedeño A., Rosso P. Overview of the 1st International Competition on Plagiarism Detection. In Stein et al. (2009), pp. 1–9. URL http://ceur-ws.org/Vol-502.

Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-based multilingual retrieval model. In Advances in Information Retrieval, 30th European Conf. on IR research, ECIR 2008, Glasgow , LNCS(4956), pp. 522–530, Springer-Verlag.

Potthast M., Barrón-Cedeño A., Eiselt A., Stein B., Rosso P. (2010a). Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler M., Harman D., and Pianta E.(Eds.), Notebook Papers of CLEF 2010 LABs and Workshops, CLEF-2010, Padua, Italy, September 22-23

Potthast M., Barrón-Cedeño A., Stein B., Rosso P. (2010b). An Evaluation Framework for Plagiarism Detection. In: Proc. of the 23rd International Conference on Computational Linguistics, COLing-2010, Beijing, China, August 23-27

Potthast M., Barrón-Cedeño A., Stein B., Rosso P. (2011). Cross-Language Plagiarism Detection. In: Languages Resources and Evaluation. Special Issue on

Plagiarism and Authorship Analysis, vol. 45, num. 1. DOI: 10.1007/s10579-009-9114-z

Scaife B. (2007). Evaluation of plagiarism detection software. Technical report, IT Consultancy,

Stamatatos E. (2009). Intrinsic plagiarism detection using character n-gram profiles. In Stein et al. (2009), pages 1–9. http://ceur-ws.org/Vol-502, pp. 38-46.

Stein B., Rosso P., Stamatatos E., Koppel M., Agirre E., Eds. SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), San Sebastian, Spain, 2009. CEUR-WS.org. http://ceur-ws.org/Vol-502.

Vallés E. (2010). Empresa 2.0: Detección de plagio y análisis de opiniones. M.Sc. thesis. Universidad Politécnica de Valencia.

Weber S. (2007). Das Google-Copy-Paste-Syndrom.Wie Netzplagiate Ausbildung und Wissen gefahrden.Telepolis.

Vers une représentation normalisée de la banque lexicale de l'IERA

Said ELHassani, Abdelfattah Hamdani

Institut d'Etudes et de Recherches pour l'Arabisation, Rabat
said.elhassani@gmail.com
fattahhamdani@gmail.com

1. Introduction

La réalisation de la plupart des applications de traitement automatique des langues naturelles (TALN) nécessite un ensemble minimal de ressources lexicales, et la réussite de ces réalisations dépend fortement des données lexicales en liaison avec le logiciel de traitement. Par conséquent, la qualité, la consistance et la normalisation des ressources lexicales est une condition préalable et importante pour le développement d'applications robustes et de large couverture.

Dans le domaine du TALN, la réalisation et l'exploitation de bases lexicales est en pleine explosion, et de nombreux travaux ont été réalisés pour développer des ressources lexicales ayant des structures différentes et répondant à des besoins différents. Ces réalisations on été accompagné par des propositions de standardisation tel que le TEI (Text Encoding Initiative) (crée par le consortium TEI en 1987), le WordNet (université de Princeton, 1993) et le LMF (Lexical Markup Framework) (Francopoulo G. et al, 2006).

Dans ce travail, nous proposons d'utiliser la norme LMF pour une représentation normalisée de la base lexicale de données de l'IERA, cette norme est une initiative récente vers des normes ISO pour la conception, la mise en œuvre et la représentation des ressources lexicales.

L'idée fondamentale de LMF est de fournir une plate-forme de spécification qui permet d'utiliser un ensemble de modules génériques (composants) qui seront combinés avec des descripteurs élémentaires (catégories de données) (Salmon-Alt S. et al, 2005). Cette spécification est prévue pour couvrir non seulement une grande variété de structures lexicales possibles, mais également un large éventail

de langues. Les principes de spécification du LMF peuvent aussi être utilisés soit comme un nouvel outil descriptif pour des ressources lexicales existantes du TAL, soit constituer une base pour la conception de nouvelles bases de données lexicales. Dans ce travail, nous essayerons d'illustrer ce premier aspect par une étude de cas sur l'utilisation de la plate-forme de spécification de LMF comme nouvel outil descriptif de la base lexicale existante à l'IERA. Pour cela, nous allons présenter les principales caractéristiques des bases lexicales de l'IERA en insistant notamment sur leurs spécificités structurelles, ensuite nous argumenterons le choix de la modélisation LMF pour la normalisation de la représentation de ces ressources en travaillant sur un échantillon du dictionnaire de la langue générale arabe/français. Nous détaillerons ensuite le modèle normalisé que nous proposons.

2. Caractéristiques lexicales de la base lexicale de l'IERA

Durant les années 80 à 90, l'IERA a entrepris la construction d'une grande banque de données lexicales divisée en deux axes, le premier axe est d'ordre terminologique et il englobe plusieurs domaines du savoir tel que l'agriculture, l'automobile, la chimie, l'électricité, l'industrie, la zoologie et autres, ces domaines de savoir sont eux aussi regroupés en trois macros disciplines : les sciences fondamentales et naturelles, les sciences appliquées et techniques et les sciences humaines et sociales. Le deuxième axe de la banque de données englobe quant à lui deux dictionnaires bilingues de la langue générale : Arabe/Français et Arabe/Anglais.

L'unité de base de la structure de la base de données est l'unité lexicale (lexème)[1]. La base se compose d'un ensemble d'équivalences de sens établies par un document (source) entre une ou plusieurs unités lexicales en langue(s) européens et une ou plusieurs unités lexicales en langue arabe qu'il fallait restituer telle se présentait dans le document (Richard N., 1987). Les relations peuvent être simples (relation un lexème arabe à un lexème européen) ou complexes (relation entre N à N).

Durant sa constitution et son évolution, la banque a connu une grande diversité au niveau de sa structure et de son contenu. Cette diversité est due non seulement à la spécificité des données traitées, mais aussi à la multitude des intervenants, aux différentes orientations stratégiques et aux contraintes techniques imposées par le

[1] Lexème : Le lexème (aussi appelé unité lexicale par le Conseil supérieur de la langue française et de nombreux grammairiens et lexicographes) est le morphème lexical d'un lemme.

degré du développement technologique dans le temps. Cette diversité peut être résumée dans les points suivants :

- La description du contenu des entrées linguistiques varie d'un dictionnaire à l'autre selon la période de sa création et ne répondait pas à un formalisme bien déterminé.
- Certaines entrées lexicales se contentent d'exemples pour définir le sens d'un mot, d'autres par un commentaire. Les informations d'ordre morphosyntaxique ne sont pas standardisées et sont plus ou moins riches d'un dictionnaire à un autre, etc.
- Les catégories des données utilisées ne sont pas unifiées au niveau de la banque et varient d'une base lexicale à une autre.

La réutilisation de ces ressources lexicales pose donc plusieurs problèmes à cause de la variation de leurs structures et de leurs descripteurs linguistiques, et l'échange de données entre les ressources lexicales est très difficile. Par ailleurs, le système de consultation accompagnant cette banque offre une recherche limitée ne dépassant pas l'aspect lexical.

Afin de faciliter les échanges de ces ressources lexicales à travers la communauté du TALN, il est important de normaliser la présentation lexicale de ces ressources pour permettre leurs fusions, leurs réutilisations et leurs interopérabilités.

3. Structures de la base lexicale de l'IERA

La structure de la banque de l'IERA a connu la succession de plusieurs versions importantes, ces versions traduisaient des ambitions des plus amples au plus réalistes et relatives tant à la qualité qu'a la quantité des données stockées. La dernière structure qui a marqué la banque de données est illustrée ci-après (Richard N., 1987) :

Le véhicule d'information dans cette structure est la relation sémantique qui existe entre deux lexèmes, cette relation est identifiée par un identifiant unique appelé "Numéro d'accession".

Les catégories d'informations distribuées dans une relation sémantique sont :

- L'unité de base appelée " Entrée " et qui est conçue comme un lexème relatif aux langues traitées (Arabe ; Français ; Anglais ; Latin).
- Des informations complémentaires :
  1. Informations sur le la relation :
     a. Source de la relation "SC"; document original d'où était tirée la relation.
     b. Domaine d'emploi "DE"; qui situe l'usage particulier des termes et donc le domaine d'application de la notion présentée par ces termes.

   c. Commentaire ou Définition "CM"

2. Informations relatives aux termes (lexèmes) dans chacune des langues qu'il traite:
   a. Catégories grammaticales (verbe ou nom ou adjectif, genre, nombre, racine, masdar, etc.), elles sont insérées à la fin du champ d'entrée lexicale entre deux #,

Ex : blanc adj., #fém. blanche#

أبيض، #ج : بيض، م : بيضاء، ج : بيضوات#

   b. Relations avec d'autres termes (Related Terms) : homonymie, synonymie, autres

L'entrée lexicale dans le dictionnaire de la langue générale de l'IERA est représenté sous une forme canonique entièrement vocalisée qu'on appelle lemme2 (un nom doit être au singulier, un verbe doit être à l'accompli avec la troisième personne du singulier etc.). Un lemme peut être formé par un mot simple ou un mot composé.

NO : (Numéro d'accession)	
SC : (Source)	مص: (مصدر)
DE : (Domaine d'emploi)	ما : (ميدان استعمال)
FR : (Entrée français) #...#	دخ: (دخلة) #...#
CM : (Commentaire)	تل: (تعليق)
EN : (Entrée anglais)	
LT : (Entrée latin)	

Ex : structure d'une relation sémantique

NO : SGH010436 **SC** : GHA **DE** : sc. nat. **FRA** : sapotillier, néflier, d'Amérique sapotier **ENG** : sapota, sapodilla plum, naseberry	ما :ع. طبيعي دخ: أخْراسَ أمْرِيكِيّ، سَبُوئَة، زُعْرُرُ أمْرِيكَةِ

[2] Lemme : mot, expression ou phrase

LAT : achras, sapota, sapota achras **CM :**	:ل‍

Ex: exemple d'une relation lexicale

Devant les limites mentionnées ci-dessus, et en vue de donner à la base lexicale de l'IERA une nouvelle vie, une modélisation de la base lexicale est nécessaire. Cette modélisation permettra de profiter de la richesse du contenu des dictionnaires de l'IERA en l'unifiant dans une structure évolutive, à partir de laquelle, il est possible de réaliser des fonctions de consultation génériques et adaptées aux besoins des utilisateurs.

4. Choix de LMF comme norme de standardisation

L'objectif de LMF est de fournir un modèle commun pour la création et l'utilisation de ressources lexicales moyennant une structure modulaire qui facilite l'interopérabilité du contenu à travers tous les aspects des ressources lexicales (Francopoulo G. et al, 2006).

La norme LMF ISO 24613 est parfaitement adaptée à notre but car elle permet :

- Une spécification des lexiques monolingues et multilingues destinés à la fois à un usage éditorial ou TALN.
- Une modélisation extensible et modulaire couvrant tous les niveaux de description linguistique (morphologie, syntaxe, sémantique).
- Une gestion séparée de la structure hiérarchique des données (méta-modèle noyau) et des descripteurs linguistiques élémentaires (catégories de données) (Salmon-Alt S. et al, 2005).
- Une certaine souplesse pour la modélisation des caractéristiques morphologiques de la langue arabe (flexionnelle et dérivationnelle) ainsi, les entrées lexicales sont représentées par la racine et le schème (Khemakhem A. et al, 2007).

5. Présentation générale de LMF

Dans sa dernière version, la norme LMF ISO-24613:2008 (qui permet de spécifier des lexiques monolingues et multilingues destinés à l'usage TALN) (Francopoulo G. et George M, 2008) a été validée convenablement pour plusieurs langues européennes, asiatiques et américaines. Cependant pour la langue arabe, des

travaux ont été réalisés dans le cadre d'un projet de coopération Tuniso-Français[3] pour confronter la norme LMF aux spécificités de la langue arabe. Ce projet a abouti à un enrichissement de la norme par les spécificités et les exigences de la langue arabe.

Un lexique LMF se présente sous forme d'un méta-modèle noyau obligatoire et un ensemble d'extensions optionnelles qui décrivent les ressources lexicales spécifiques en réutilisant les composants du noyau (Francopoulo G. et al, 2006). Le méta-modèle noyau forme une structure hiérarchique des classes UML qui spécifie les notions de lexique, de l'entée lexicale, de forme et de sens. LMF fournit un mécanisme permettant de spécifier le contenu des classes du méta-modèle noyau à l'aide de descripteurs élémentaires sous forme de couples "Attribut-Valeur" définis par une autre norme ISO 12620, appelée catégories de données (RCD) (Romary L. et al, 2003) consultable et éditable en ligne (http://syntax.inist.fr.). Les catégories de données reflètent les concepts de base linguistique, tels que /partOfSpeech/, /Genre/, /Nombre/ et ils sont stockés et gérés indépendamment de la structure hiérarchique du modèle de données.

6. Modèle normalisé des dictionnaires de la langue générale de l'IERA

En se référant à la dernière révision de LMF v.16 (Francopoulo G. et George M., 2008), notre lexique sera limité à la représentation de niveau morphologique qui nous paraît nécessaire pour la plupart des applications TAL.

Nous avons retenu cinq classes pour la modélisation du noyau à savoir : Database, Lexicon, Lexicon Information, Lexical Entry, Form dont les deux sous-classes de spécification de Form: LemmatisedForm et InflectedForm. Toutes ces classes sont représentées dans la Fig.1.

Pour le codage de l'information flexionnelle, nous avons adopté délibérément une perspective extensionnelle, c.-à-d. une description de l'ensemble des formes (LemmatisedForm et InflectedForm) pour une entrée lexicale donnée. Ainsi nous allons utiliser un conjugueur arabe (développé à l'IERA) pour récupérer les mots fléchis de chaque LemmatisedForm des entrées lexicales du modèle.

---

[3] Entre le laboratoire MIRACL de l'université de Sfax en Tunisie, et le laboratoire LORIA/INRIA en France.

En langue arabe, la majorité des mots (verbe et noms dérivables) sont identifiées par une racine et un schème, ceci implique la spécification de la classe LexicalEntry par les catégories de données /root/, /schème/ et /pos/, dont les valeurs sont prises du RCD. La classe Form représente les différentes variations orthographiques et phonologiques de la classe LexicalEntry avec des spécifications grammaticales. La combinaison d'une racine et d'un schème peut générer une ou plusieurs LemmatisedForm qui représentent le lemme de LexicalEntry. Un lemme peut avoir plusieurs InflectedForm qui représentent les formes fléchies correspondant à une forme d'occurrence d'une LemmatisedForm Fig. 2.

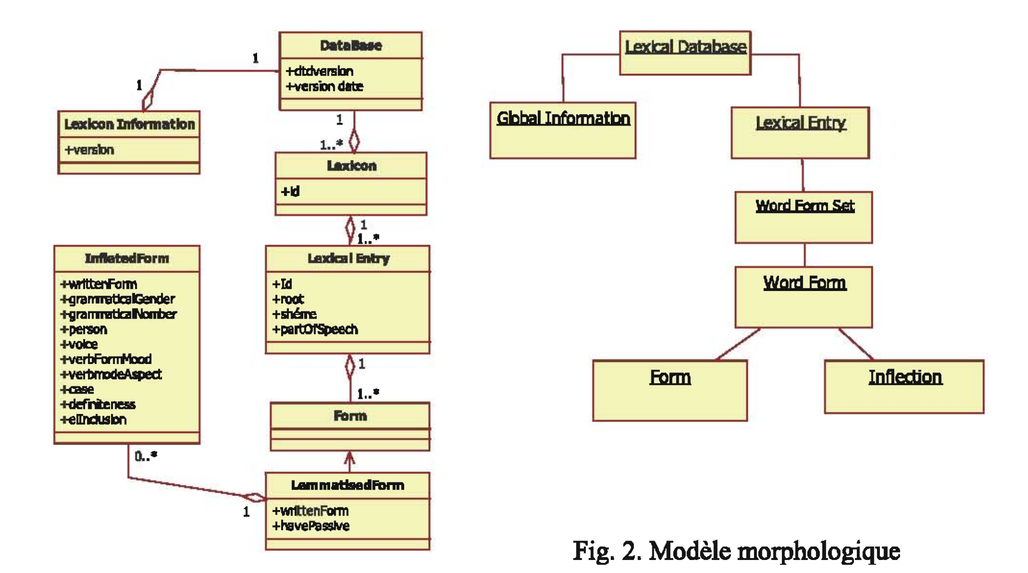


Fig. 1. Le modèle noyau avec l'extension morphologique (extensionnelle)

Fig. 2. Modèle morphologique

7. Les catégories de données du modèle morphologique

Le lexique de la langue arabe comprend trois catégories de mots : verbe, nom et particule. Pour notre modèle il s'impose maintenant de décider quelles sont les propriétés flexionnelles qui sont appropriées pour chaque catégorie de mots. Dans le cadre de ce papier, nous nous intéresserons seulement aux catégories morpho-syntaxiques des noms et des verbes.

Cas des Noms

Les noms arabes ont plusieurs sous-catégories qui peuvent être variables ou invariables, généralement les noms variables ont plusieurs formes fléchies qui sont associées aux catégories de données (Voir Tab1). Les noms arabes portent des informations grammaticale sur le genre, mais concernant le nombre, la particularité de l'arabe est d'avoir un système à quatre valeurs: singulier, dual, pluriel et pluriel brisé. Les noms arabes sont également soumis à une variation de cas: nominatif, accusatif et génitif. En outre, ils sont définis de trois façons : soit avec l'article ال, soit par un syntagme nominal (الإضافة) ou soit par des pronoms personnels dans une structure possessive (كتابي). Les noms arabes sont aussi soit définissable par ال , soit non définissable par ال quand il s'agit de noms propres ou de noms indéterminés par la désinence " Tanwin".

Data Category Identifier	Conceptual Range
/wordForm/	orthographe de la forme fléchie
/grammaticalGender/	{/masculine/, feminne/, /neutre/}
/grammaticalNumber/	{/singular/, /dual/, /plural/, /plural broker/}
/grammaticalCase/	{/nominative/, /accusative/, /genetive/}
/grammaticalDefiniteness/	{/indefinite/, /definite/}
/elInclusion/	{/yes/, /no/}

Tab1. Catégories de données des noms

Cas des verbes

Les verbes arabes sont soumis à un système de variation flexionnelle en associant, à chaque forme fléchie ou combinaison de traits morphologiques d'un verbe, les catégories de données mentionnées dans le tab2. En outre, la combinaison de ces caractéristiques est conditionnée par des contraintes de cooccurrence particulière: le mode ne s'appliquant que pour l'inaccompli, le genre s'applique seulement avec $2^{ème}$ et $3^{ème}$ personne et la voix passive est incompatible avec le mode impératif. En plus de ces catégories de données, les verbes arabes varient également en termes de nombre, personne et genre grammatical.

Data Category Identifier	Conceptual Range
/wordForm/	orthographe de la forme fléchie
/grammaticalNumber/	{/singular/, /dual/, /plural/}
/grammaticalGender/	{/masculine/, feminine/, /neutre/}
/grammaticalPerson/	{/firstPerson/, /secondPerson/, /thirdPerson/}
/grammaticalAspect/	{/Accomplished/, /Unaccomplished/, / Imperative /}
/grammaticalVoice/	{/active/, /passive/}
/grammaticalMood/	{/indicative/, /subjunctive/, /jussive/}

Tab2. Catégories de données d'inflexion de verbe

Généralement le nombre de traits morphologiques peut varier d'une forme fléchie à une autre parce qu'il y a des traits morphologiques dont la présence dépend d'un autre trait : par exemple le genre est absent avec la première personne (/firstPerson/).

8. Mise en œuvre du modèle morphologique

Dans une première étape, nous limiterons la mise en œuvre de notre modèle au traitement des verbes arabes, et dans une étape ultérieure, elle sera généralisée aux noms.

En suivant la représentation extensionnelle du modèle, et en utilisant le conjugueur des verbes développé à l'IERA, nous allons pouvoir récupérer toutes les formes fléchies des verbes arabes Fig. 3.

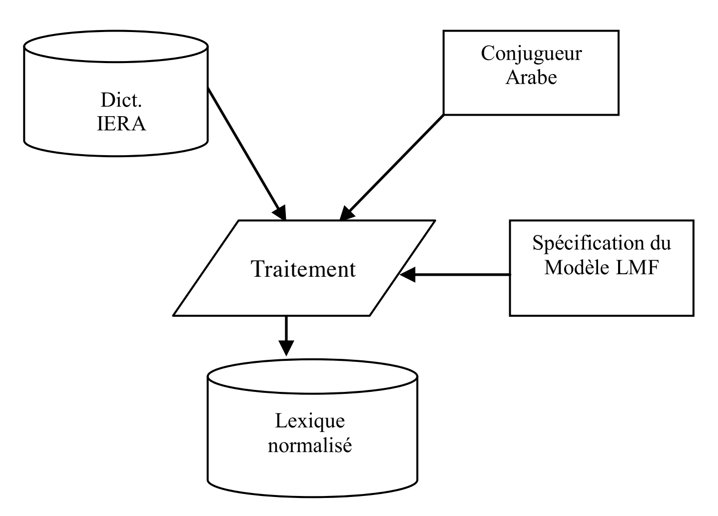


Fig .3. Architecture de construction du lexique normalisé

Concernant les noms arabes, nous utiliserons l'analyseur morphologique des mots arabes (élaboré aussi à l'IERA) pour extraire toutes les informations nécessaires à la constitution du lexique. Ici, une intervention humaine de vérification et de validation est nécessaire vue que l'analyseur ne pourra pas analyser la totalité des noms arabes, néanmoins, un grand nombre de noms pourront être analysés automatiquement. Ces formes fléchies sont accompagnées de leurs catégories de données mentionnées dans le tab. 2.

9. Conclusion

Dans ce papier nous avons proposé un modèle conforme à la norme LMF ISO 24613 qui permet de normaliser la représentation des ressources lexicales de l'IERA en construisant un lexique plein forme à usage TALN à partir du dictionnaire de la langue générale. Ce lexique est ouvert à toute possibilité d'extension.

Actuellement, la mise en œuvre du lexique se limitera dans un premier temps aux verbes arabes. Dans une seconde étape, elle sera généralisée aux noms arabes.

Durant ce travail, nous avons essayé de focaliser notre attention sur la modélisation des bases lexicales de l'IERA selon des normes internationales. Cependant, et concernant l'exhaustivité des trais morphologiques et linguistiques, le modèle est certes extensible, mais l'accomplissement de ce travail nous amènera à travailler d'avantage et en étroite collaboration avec des linguistes afin d'enrichir la structure du lexique par des catégories de données appropriés.

Références

(Blachère et al, 1975) Blachère R., Gaudefroy-Demombynes M., "Grammaire de l'arabe classique", Edition Maisonneuvre-Larose, Paris, 1975.

(Francopoulo G., 2004) Francopoulo G., "Proposition de normes des lexiques pour le traitement automatique de la langue", INRIA/LORIA-ACTION SYNTAXE, Version-1.10 13 mai 2004.

(Francopoulo G. et al, 2006) Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. Lexical Markup Framework (LMF), LREC 2006, Genoa.

(Francopoulo G. et George M., 2008) Francopoulo G., George M. (2008). ISO/TC 37/SC 4 N453 Rev.16. Language resource management – Lexical markup framework (LMF).

(Khemakhem A. et al, 2007) Khemkhem A., Gargouri B., Abdelwahed A., Francopoulo G. (2007). Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF - ISO 24613. Traitement Automatique des Langues Naturelles : du 5 au 8 juin 2007 à Toulouse.

(Romary, 2003) Romary L., "Action nationale INRIA Syntax (Décembre 2001 – Décembre 2003)", INRIA, 2003.

(Romary et al, 2003) Romary L., Wright S., Farrar S., Gillam L.,ISO TC 37/SC4 N055, Language resource management - Implementing a data category registry within ISO TC37, 2003.

(Romary et al, 2004) Romary L., Salmon-Alt S., Francopoulo G., "Standards going concrete : from LMF to Morphalou", Workshop on Electronic Dictionaries, Coling2004,Geneva, Switzerland, 2004.

(Richard N. 1987) Richard N., "Arabisation et technologie" Edition de l'IERA, Rabat, 1987.

(Salmon-Alt S. et al, 2005) Salmon-Alt S., Akrout A., Romary L., "Proposals for a normalized representation of Standard Arabic full form lexica", Second International Conference on Machine Intelligence (ACIDCA-ICMI 2005).

Proposition pour la création d'un groupe TEI berbère fédérant la mise en chantier d'un sous-ensemble de Guidelines spécifiques assurant la qualité d'interopérabilité des ressources linguistiques amazighes

Henri Hudrisier

Laboratoire Paragraphe et LEDEN
Henri.hudrisier@wanadoo.fr

En tant que chercheur non-berbérophone, mais cependant associé à de nombreuses occasions à des travaux de réflexion initiés par des institutions dédiées à la culture berbère, j'ai pu bénéficier d'une certaine neutralité d'observation en qui concerne la mise en place des projets d'environnements techniques numériques berbères.

Comme dans de très nombreux autres chantiers de rassemblement de ressources patrimoniales et linguistiques, le problème est certes technique (il est très bien et très exhaustivement décrit dans les thématiques de l'appel à communications de ce 4e workshop) mais ce qui manque toujours le plus c'est une réelle dynamique de fédération mondialisée des intervenants et des institutions. Cette situation est commune à tous les projets[1] et elle est trop souvent sous-estimée.

De ce point de vue, le travail qui me semble urgent à mettre en chantier consiste à fédérer les savoirs et les savoir-faire des intervenants du « numérique berbère ». Pour réaliser ce projet, il manque un « super forum ou super Wiki numérique »

[1] Parce qu'on achoppe inévitablement sur la légitime identité des personnalités, des institutions et des Etats qui ne se dissolvent pas dans l'espace numérique mondial. On achoppe aussi très souvent sur le retour sur investissement des différents partenaires potentiels des projets de « réseau numérique ». De ce fait les réseaux numériques des multinationales à visée commerciale se mettent en place beaucoup plus facilement que des projets similaires à finalité culturelle dans lesquels les partenariats financiers ne sont jamais explicites.

susceptible d'assurer à la fois leur fédération en tant que groupe d'intervenants apportant chacun ses contraintes propres[2] mais aussi à même de leur offrir un environnement numérique qui soit aussi un « espace numérique de travail » et une « boîte à outils » qui leur soient communs. Il faut aussi que cet espace de travail et cette boîte à outils soient susceptibles de se paramétrer en fonction des besoins propres à tel ou tel ou tel métier, discipline scientifique ou spécificité des corpus linguistiques[3].

De mon point de vue, la TEI est « la » solution qui fait aujourd'hui l'unanimité pour répondre précisément à cette double contrainte : fédérer les chercheurs en ressources linguistiques et leur offrir un environnement de travail répondant à leurs diverses exigences et susceptible en même temps de mettre en synergie cette diversité de besoins pour ne pas en faire autant d'obstacles à un travail qui se doit de rester une mise en synergie.

1. La TEI : un cadre tant technique que collaboratif adapté à la gestion numérique des documents à haute valeur ajoutée

La TEI que l'on pourrait traduire par « groupe d'initiative pour l'encodage normalisé des textes » est un standard de balisage, de notation et d'échanges de corpus de documents électroniques fondé sur l'utilisation systématique de langages balisés (Markup Language comme le SGML et maintenant le XML) spécialement aménagés pour permettre la pose virtuelle de balises (des signets) tant sémantiques, que structurels ou référentiels. L'originalité de la communauté des chercheurs (à l'origine en majorité des spécialistes des études littéraires associés à des bibliothécaires et des informaticiens spécialistes de la numérisation des textes), tient à ce qu'ils ont très tôt compris que le traitement sémantique des documents numériques dépend, certes des nouvelles opportunités de ces langages balisés, mais aussi et d'abord, de la capacité des chercheurs à s'entendre, en consensus pour définir les fondamentaux de leur discipline de recherche et créer ainsi des TEI

[2] À leur métier, à leur discipline scientifique, à leur parler berbère.

[3] Spécifique par la langue (tel ou tel parler berbère ou telle langue autre traitant de sujet berbère), spécifique par le média (TV, radio, ressources écrites manuscrites ou imprimées, ressources numériques) ou le genre (corpus oraux ou écrits, dictionnaires, littérature savante de jeunesse ou grand public, contes, poèmes ou chansons etc…)

spécialisées à même de faciliter l'échange de leurs résultats. Ont été ainsi créées la *TEIverse* pour les études poétiques, la *TEIdrama* pour les études théâtrales et ma proposition serait que nous puissions fonder une *TEIberbère* ce qui n'a rien d'une utopie.

La fondation de la TEI remonte à une conférence qui s'est réunie au Vassar Collège à Poughkeepsie (NY-USA) en novembre 1987. Les discussions, dans un groupe d'une trentaine de chercheurs venant des domaines de la bibliothéconomie, des sciences humaines, de la littérature et de la recherche informatique[4] ont abouti à des recommandations pour définir un système commun d'encodage des documents textuels : ce colloque du Vassar Collège est à l'origine du « groupe d'initiative TEI[5] » qui fut créé officiellement en 1988 par trois associations professionnelles ayant des activités de recherche dans le domaine du traitement de textes par ordinateur:

- l'*Association for Computational Linguistics (ACL[6]),*
- l'*Association for Literary and Linguistic Computing (ALLC[7]),*

---

[4] Ils avaient participé à la recherche développement de SGML *(Standard Generalized Markup Language),* le langage à balises *(Markup Language)* historique qui a précédé HTML *(Hyper Text Markup Language)* puis XML *(Extended Markup Language).*

[5] Il faut bien comprendre le mot initiative dans son sens anglais qui peut être à la fois l'action mais aussi un groupe de pression, un groupe de travail créé pour faire avancer une action.

[6] L'Association pour la Linguistique Informatique est une société savante et professionnelle internationale pour tous ceux qui s'intéressent aux questions posées par l'informatisation du langage naturel. L'adhésion inclut l'ACL le journal trimestriel, la Linguistique Informatique, l'abonnement à l'édition résumée des conférences et la participation dans des groupes de travail et d'études ACL. Le journal ACL : Linguistique Informatique est aujourd'hui le forum principal pour la recherche sur la linguistique informatique et le traitement de langage naturel. Depuis 1988, le journal a été publié pour l'ACL, les presses du MIT lui assurent une base de distribution mondiale.

[7] L'Association pour la Linguistique Computationelle Littéraire a été fondée en 1973 dans le but de soutenir l'application de l'informatique dans l'étude de la langue et de la littérature. Les intérêts des membres de l'Association se sont nécessairement élargis, (avec le progrès de l'ingénierie du langage) pour englober non seulement l'analyse de textes et des corpus de langue, mais répondre aussi au traitement des éditions électroniques. L'adhésion à

- L'Association *for Computing and the Humanities (ACH*[8]*)*.

La TEI a été financée au départ par :

- *l'US National Edowment for the Humanities*,
- la Communauté européenne (DG13),
- le *Social Science and Humanities Research Council* du Canada,
- la Fondation Andrew W. Mellon.

Notons que des informaticiens pionniers fondateurs de la TEI ont été directement à l'origine de certains développements fondamentaux de XML[9]*(Extended Markup Language)*. C'est la raison pour laquelle la TEI qui a été créée à ses débuts sur un substrat logiciel SGML a maintenant totalement migré sur XML.

L'originalité de cette communauté de chercheurs (volontariste dans son amalgame multidisciplinaire), tient à ce qu'ils ont très tôt compris que le traitement sémantique des documents numériques découlait certes des nouvelles possibilités offertes par les langages balisés qui étaient en pleine émergence avec le SGML *(Standard Generalized Markup Language)*, mais aussi, et en premier, dépendait de la capacité des chercheurs à s'entendre en consensus pour définir les fondamentaux de leurs disciplines de recherche et créer ainsi des TEI spécialisées, à même de

---

l'ALLC'S est ouverte à tous les pays du monde et à toutes les disciplines que l'on qualifie sous les termes génériques de disciplines littéraires (chercheurs ou étudiants). L'Association édite un journal : *Literary and Linguistic Computing*, publié par *Oxford University Press*. Chaque année l'association organise une conférence plénière, en collaboration avec *l'Association for Computers and the Humanities*. Les conférences plénières se situent alternativement en Europe et en Amérique du Nord.

[8] L'Association pour l'usage de l'informatique dans les lettres et autres sciences humaines qualifiables du terme générique : Humanités. Depuis sa fondation, ACH a été la société professionnelle internationale de référence pour la recherche assistée par ordinateur pour la littérature et les études de langue, l'histoire, la philosophie et d'autres disciplines d'humanités. L'ACH est particulièrement impliquée dans les recherches sur la manipulation et l'analyse de corpus textuels. L'ACH s'attache particulièrement à disséminer des informations et des logiciels parmi les membres de ses groupes de travail. Elle encourage aussi le développement et la dissémination de ressources textuelles et linguistiques significatives.

[9] Lou Burnard et C.M. Sperberg Mc Queen.

faciliter la numérisation de leurs recherches et de l'échange de leurs résultats. Les fondateurs de la TEI souhaitaient un système de balisage et un format communs standardisés facilitant le traitement par ordinateur, l'échange et le partage des textes numérisés.

Citons deux de ses fondateurs (Ide N. & all 1996) relatant les exigences techniques et scientifiques de sa fondation.

« À l'époque, l'énorme variété des formats de codage et de représentation des textes (à peu près tous mutuellement incompatibles) était perçue comme un obstacle majeur à l'échange des données et à la recherche. Les chercheurs présents à Vassar sont tombés d'accord sur la nécessité de travailler à la définition d'un nouveau format de codage des textes électroniques et en ont posé les principes de base. Le nouveau format devait:

- être aussi complet que possible,
- être simple, clair et concret,
- être facile à utiliser sans logiciel particulier,
- être rigoureusement défini,
- permettre un traitement efficace,
- être ouvert à des extensions définies par les utilisateurs, être compatible avec les standards existants ou en développement.

[...] De nombreux chercheurs à travers le monde ont travaillé regroupés dans des comités traitant chacun d'un thème précis. L'ensemble a été coordonné par un Comité de Pilotage (présidé successivement par Nancy Ide, Don Walker, Susan Hockey et David Barnard) et deux éditeurs (Michael Sperberg-McQueen et Lou Burnard).

En mai 1994, le travail effectué par les différents comités a été publié sous forme de *Guidelines for Electronic Text Encoding and Interchange* (« Recommandations pour le codage et l'échange des textes informatisés »), aussi connues sous le nom de « TEI P3 ». Ces *Recommandations* proposent un ensemble de conventions de codage utilisable dans une grande variété d'applications : publication électronique, analyse littéraire et historique, lexicographie, traitement automatique des langues, recherche documentaire, hypertexte, etc. C'est aujourd'hui un consortium académique international, créé en 1987, dans le but de développer les recommandations pour le codage et l'échange de données linguistiques et

littéraires. En mai 1994, le travail effectué par les différents comités a été publié sous forme de « Recommandations pour le codage et l'échange des textes informatisés » (*Guidelines for the Encoding and Interchange of Machine-Readable Texts*), aussi connues sous le nom de TEI P3, reposant sur les DTD du SGML. »

Ces directives proposent un ensemble de conventions de codage utilisables dans une grande variété d'applications : publication électronique, analyse littéraire et historique, lexicographie, traitement automatique des langues, recherche documentaire, hypertexte, *etc*. Les directives concernent les textes écrits ou parlés, sans restriction de langue, de période, de genre ou de contenu et répondent aux besoins fondamentaux de nombreux d'utilisateurs: lexicographes, linguistes, philologues, bibliothécaires, et de manière générale, de tous ceux qui sont concernés par l'archivage et l'accès à des documents électroniques.

Trois aspects du codage des textes sont particulièrement mis en avant par la TEI :

- **documentation de textes** : les documents TEI doivent fournir obligatoirement les informations bibliographiques sur le texte lui-même et son codage. Ces informations sont balisées dans la partie en-tête « TEIheader » se trouvant au début de chaque document codé en TEI.Ceci est particulièrement important parce que grâce à ses différents « desks (ou zones)[10] », le TEIheader permet de documenter beaucoup mieux qu'une fiche bibliographique classique les différents niveaux et versus d'un document numérique : distinguer la source (document primaire) de ses états numériques, pouvoir documenter les auteurs, les dates et institutions de la numérisation ou ensuite des différents balisages.

- **représentation de textes** : la TEI propose un système de balises pour coder la description de structure logique de différents types de documents (textes écrits ou parlés, prose littéraire, poésie, théâtre, dictionnaires, données terminologiques, hypermédias etc.)

- **analyse et interprétation de textes** : les directives de la TEI contiennent des jeux de balises pour le codage des références croisées ou des index dans les textes, des analyses linguistiques et des informations concernant l'étude littéraire.

Soulignons aussi l'importance de la communauté des « activistes et des fondateurs de la TEI » dans la recherche, le développement, puis la mise en œuvre effective de projet de gestion des documents numériques à haute valeur ajoutée. On ignore souvent que plusieurs informaticiens fondateurs de la TEI ont été dans l'équipe qui

---

[10] <fileDesc> <encodingDesc <profileDesc> <revisionDesc> cf infra.

a développé XML. Ce contexte particulièrement riche en expertise nous permet d'envisager les conditions de constitution d'un collège international de chercheurs à même d'amorcer, puis de finaliser la mise au point d'un modèle de déploiement numérique de ressources numériques structurées par *TEIberbère*.

2. Les enjeux scientifiques de la TEI

La TEI a été mise au point pour que des chercheurs, au début, principalement des chercheurs en sciences humaines, puissent non seulement échanger des corpus de textes, mais aussi disposer en commun d'un système de balisage et d'annotations normalisées. SGML, comme on le sait, est à l'origine un balisage issu de l'organisation des textes nécessaires aux éditeurs. Le coeur de la TEI reprend les éléments d'analyse nécessaires pour décrire la structuration fonctionnelle d'un texte (titre, avertissement, préface, corps du texte décomposé en chapitres et sous chapitres, index, table des matières, etc.) initié avec SGML. Il a été très significativement augmenté pour constituer ce que nous sommes convenus d'appeler le « noyau TEI ».

Selon la discipline à laquelle appartient un chercheur utilisant la TEI, il lui sera ensuite commode d'utiliser, au-delà de ce noyau, les éléments de niveau disciplinaire qu'il jugera utile à sa recherche. L'aménagement de textes par des chercheurs, selon la norme TEI, permet dès lors, que des chercheurs en littérature, en histoire, en ethnologie, etc. puisse ainsi, chacun dans leur discipline propre, (et même hors de leur discipline), procéder à des échanges de corpus comprenant aussi bien les textes que leurs annotations conceptuelles.

On comprend ainsi que (contrairement à ce que pensent parfois des professionnels de la documentation ou des gestionnaires de gros corpus de documents), la TEI est beaucoup plus qu'un simple format d'échange de gros corpus de textes. C'est aussi un vaste forum d'échange et d'accumulation des apports conceptuels d'autres chercheurs en sciences humaines. Avant la TEI, cette transmission ne pouvait se réaliser que par la lecture et « l'assimilation intellectuelle individuelle » des articles et ouvrages, suivies d'une reprise personnelle non instrumentalisable[11] des éléments du corpus selon les résultats transmis par ces articles et ces ouvrages. La TEI ne dispense pas de lire nos collègues, bien au contraire, mais elle nous permet, comme en sciences exactes, de disposer directement et de façon normalisée

[11] Les informaticiens disent non calculable, non « computarisable ».

numérique et immédiatement utilisable des textes « traités » selon les hypothèses d'autrui.

C'est cela qui change tout. La TEI permet ainsi de mettre en chantier et de façon très facilement collaborative de vastes projets linguistiques, d'analyse littéraire, philologique, d'analyse comparée multilingue, etc… Le système de balisage étant défini à la fois entre tous de façon multidisciplinaire, puis, ensuite au niveau de chaque discipline, les corpus peuvent ainsi être « augmentés » en passant de l'un à l'autre. On peut aussi (et c'est très productif pour les travaux en TAL), bénéficier d'autres chercheurs travaillant sur d'autres langues et bénéficier de tout ou partie de leurs conventions de balisage.

<u>Le contexte matériel de la définition de ces fonctions de balisages</u>

On l'a déjà souligné, il existe une synergie et une similitude entre ce qui peut être fait en XML et ce que permet la TEI. Il faut cependant souligner une importante différence :

- XML (qui après SGML est l'outil de balisage structurel, référentiel et sémantique par excellence) ne peut fonctionner comme un moyen d'échange sémantique que si les partenaires de ce réseau d'échange partagent la même sémantique des balises. Pour ce qui est des balises structurelles et référentielles (renvois bibliographiques par exemple) la sémantique est assez largement commune et elle recoupe la sémantique de balisage de HTML qui s'impose maintenant à tous. En revanche, pour ce qui est de la sémantique des « balises sémantiques », elle est, par construction, ouverte sur l'infini des possibles.
- Les balises TEI prennent dès lors toute leur valeur. L'initiative TEI fonctionne comme une « fédération, voire une confédération » de collèges de chercheurs qui partagent dans chacun des collèges spécialisés des sémantiques selon la logique de cette hiérarchie de fédérations et confédérations.

Ainsi, tous les membres de la TEI partagent un noyau sémantique, chaque discipline *TEIdrama* (théâtre), *TEIverse*, etc. partage une sémantique augmentée de la spécialité ; puis, au-delà, des sous-groupes de chercheurs (études poétiques élisabéthaines par exemple) peuvent définir leurs propres jeux de balises. Ce sont des dialectes, en quelque sorte, sauf que toutes ces sous-sémantiques peuvent parfaitement être intégrées dans un même univers sémantique parfaitement cohérent et interopérable dans son ensemble.

Les fondateurs de la TEI posaient comme hypothèse majeure qu'il était possible d'utiliser la démarche de structuration par balisage[12] pour analyser des textes et noter de façon normalisée les éléments décrits par ce balisage. Ce balisage s'organise selon deux types d'éléments :

- le noyau : (ce sont des balises et des éléments communs à toutes disciplines). Par exemple, la structure en divisions et paragraphes, la description documentaire du contenu, etc.
- les balises et éléments propres à des disciplines qui permettent de travailler sur la prose, la parole, le théâtre, la poésie, les dictionnaires, l'histoire...

Fonctionnellement, le balisage TEI s'organise aussi selon deux champs complémentaires mais distincts :

- l'en-tête (*header*) qui constitue une codification non seulement de la source du document (un livre édité sur papier par exemple) mais de sa transcription numérique : personne et institution, responsables de la transcription, format de transcription, date, mode de disponibilité, versions et mise à jour, codification selon des modes de description qui peuvent être en partie automatisés pour transformer des données bibliographiques traditionnelles, etc.
- le balisage proprement dit du document. Celui-ci peut se contenter d'être relativement léger et strictement formel, ce qui permet d'échanger des références ou des corpus. Dans d'autres cas, la TEI peut devenir le support de descriptions beaucoup plus fines dans lesquelles on liera le fond et la forme du document (les études littéraires théâtrales ou poétiques sont un bon exemple de ce type de traitement.)

Plus techniquement il existe 3 ensembles de balises :

- **Un ensemble de balises obligatoires (core tag sets)**

Cet ensemble a deux composantes:

1. l'ensemble des éléments et des attributs requis pour tous les genres de documents. Par conséquent cet ensemble est obligatoire.
2. un en-tête qui peut être comparé à une page de titre électronique (*TEI header*).

- **Un ensemble de balises de base (base tag set)**

---

[12] A l'époque le SGML et maintenant, bien sûr ,le XML.

L'utilisateur doit ici choisir parmi les six ensembles définis qui représentent autant de catégories de textes: prose, poésie, théâtre, transcription du discours (*transcribed speech*), dictionnaire et informations terminologiques. Les ensembles de balises de base définissent les types de documents. Par exemple, *TEIdictionnaries* est l'ensemble qui contient la déclaration des éléments nécessaires au balisage d'un dictionnaire. Idéalement, un seul ensemble de balises propres à la discipline est nécessaire pour l'encodage d'un genre spécifique[13].

- **Ensembles de balises additionnelles (*additional tag sets*)**

Ces balises permettent de répondre à des besoins particuliers. L'utilisation de ces balises est compatible avec tous les ensembles de base. C'est ici que viendrait s'insérer à terme *TEIberbère*.

En 1994 la TEI a publié « Les recommandations de la TEI », (*TEI guidelines*) dont elle propose une « version allégée : « la TEI lite », conçue pour donner accès à un ensemble plus facile à appréhender permettant ensuite aux chercheurs de s'approprier plus facilement la totalité du *TEI guidelines*. L'ensemble de ces recommandations a été traduit en français par François Role dans le n° 24 Spécial TEI des Cahiers Gutenberg (actuellement disponible en ligne)[14].

3. La mise en synergie des chercheurs TEI

Ce point est important. Il est en effet indispensable que l'aménagement d'un espace normalisé de travail puisse être paramétrable par les chercheurs qui le souhaitent et cependant, que les normes de balisage des documents ne se multiplient pas de façon exponentielle. C'est précisément un des objectifs primordiaux de la TEI : articuler à plusieurs niveaux le processus de normalisation et de consensus : créer des consensus de description sémantique ou de structuration des documents ou des corpus qui soient communs à tous, puis sur ce premier étage (proprement normatif) laisser chaque chercheur, chaque aménageur de fonds de documents paramétrer ses balises quand c'est indispensable (si possible en rajoutant des attributs aux balises et en évitant d'en créer de nouvelles).

Ce processus d'articulation et d'enchâssement des modèles peut, et même doit

---

[13] En fait cette règle vaut plus pour l'analyse préalable de constitution d'un domaine (définir les balises spécifiques) que pour l'usage.

[14] < www.gutenberg.eu.org/publications/cahiers/50-cahiers24.html >

avoir plusieurs niveaux, entassés ou même quelquefois parallèles[15]. Par contre s'il s'avère qu'un mode de structuration ou une nouvelle catégorie sémantique se faisait jour, il est important de concevoir une (ou des) nouvelle(s) balise(s) ou mode de structuration, puis de la tester expérimentalement en local, mais il faut se garder de mettre en ligne des corpus utilisant ces nouvelles balises sans les soumettre préalablement à l'approbation en consensus de la communauté de la sous catégorie TEI concernée qui envisagera éventuellement d'intégrer ces modes de balisage innovants dans les mises à jour périodiques des Guidelines de la TEI. C'est cette dynamique d'aménagement collégial d'un sous domaine de la TEI qui répond précisément à cette double contrainte : normaliser techniquement un environnement numérique berbère et malgré tout ne pas entraver les besoins spécifiques des chercheurs.

La communauté TEI est maintenant riche de plus de 20 ans d'expérience correspondant à des mises en synergies similaires. S'associer au monde de la TEI c'est s'assurer aussi l'aide ou l'expertise de chercheurs non-berbérophones ayant déjà eu l'expérience de situations patrimoniales similaires dans d'autres domaines et d'autres langues.

Fort de cela, nous proposons l'instauration d'un groupe « TEI Berbère » avec un spectre de fédération de recherche relativement large et multidisciplinaire qui correspondra de ce fait à des « groupes de travail TEI Berbère » correspondant à des différentes tâches:

- Mise au point urgente d'un « TEIheader » propre aux divers patrimoines berbères ou ayant le berbère comme objet : c'est la première étape d'un travail permettant d'assurer une gestion de ces patrimoines en tant que bibliothèque numérique
- Analyse littéraire de ressources tant orales qu'écrites (à la fois pour des ressources directement berbères mais aussi de la littérature scientifique dont la culture berbère est l'objet) : c'est l'objet premier de la TEI dès sa fondation.
- Balisages multilingues : la TEI est particulièrement adaptée à la structuration de corpus multilingues et notamment à leur alignement parallèle.

---

[15] Par exemple dans la TEI, il existe une DTD TEIverse qui constitue un métamodèle général pour le balisage de la poésie. Par contre à l'évidence la poétique anglaise, française, latine ou berbère ne fonctionne pas selon les mêmes règles et structures ce qui implique que les communautés d'études considérées puissent décrire et baliser ces différents corpus d'étude en définissant communauté par communauté leurs modèles spécifiques.

- Rassemblement cohérent des ressources terminologiques : la DTD TEI « TEIdictionnary » et le groupe d'experts qui y est associé est particulièrement adapté. Un de ses membres actifs Laurent Romary.
- La description des caractéristiques de traits dans toutes les langues (travaux menés en consortium précisément par l'ISO TC37 SC4 & la TEI), qui mérite d'être mis en chantier dans le cadre spécifique d'une ou préférentiellement de plusieurs langues berbères.

Notons que le point 2 demande à être ultérieurement subdivisé en de nombreuses sous-thématiques : TEIdramaBerbère, TEIverseBerbère. TEItranscribedSpeachBerbere[16]…

Il est aussi vraisemblable que les chercheurs impliqués dans ces différentes sous-thématiques devront créer des sous-modèles plus ou moins spécialisés ou adaptés selon la langue berbère qu'ils étudient. Et cependant il est fondamental de rester dans une limite de granularité raisonnable. La spécification excessive des modèles TEI peut entraîner une babélisation de la recherche. Avant de créer des infinités de modes de modélisation, il importe d'explorer les balisages déjà expérimentés dans la communauté TEI. Il importe d'explorer l'usage qui pourrait être fait en

---

[16] Nous donnons ici à titre d'exemple une partie du TEIheader d'un enregistrement oral dont on précise la source vidéo d'origine

```
<recordingStmt>
 <recording type="video">
  <p>U-matic recording made by college audio-visual department staff,
    available as PAL-standard VHS transfer or sound-only casssette</p>
 </recording>
</recordingStmt>
<recordingStmt>
 <recording type="audio" dur="P30M">
  <respStmt>
   <resp>Location recording by</resp>
   <name>Sound Services Ltd.</name>
  </respStmt>
  <equipment>
   <p>Multiple close microphones mixed down to stereo Digital
     Audio Tape, standard play, 44.1 KHz sampling frequency</p>
  </equipment>
  <date>12 Jan 1987</date>
 </recording>
</recordingStmt>
<recordingStmt>
 <recording type="audio" dur="P15M" xml:id="rec-3001">
  <date>14 Feb 2001</date>.
```

réutilisant des balises conçues pour d'autres langues puis de ne les caractériser qu'au seul niveau d'un attribut : procédure qui permet dès lors de ne pas multiplier à l'infini le nombre des balises des guidelines permettant ainsi que cela reste un environnement et un outil appréhendables et compréhensibles.

4. La TEI et la diversité linguistique

Citons Nguyen Thi Minh Huyen (Nguyen T. 2006) pour ce qui est précisément du traitement automatique des langues dans le contexte d'une langue qui n'est pas précisément de grande diffusion : le vietnamien. Il souligne, l'importance du consortium TEI qui s'est lui-même associé en consortium avec l'ISO TC37 (comité technique dédié à la normalisation de la terminologie et des ressources linguistiques) :

« Avec la maturité de développement des standards dans le domaine de langues (TEI, EAGLES/ISLE, LISA[17]), l'ISO a validé en août 2002 la création d'un sous-comité TC37/SC4[18] entièrement dédié à la normalisation de la gestion des ressources linguistiques, sous la présidence de Laurent Romary. L'ISOTC37/SC4 a pour but de développer des principes et méthodes pour la création, l'encodage, le traitement et la gestion des ressources langagières comme des corpus écrits ou oraux, des lexiques, des schémas de classification. Les centres d'intérêts sont : la modélisation de données, le balisage, l'échange de données et l'évaluation des ressources langagières (à l'exception) des terminologies (traitées précédemment par d'autres sous-comités du TC 37) » .

Nguyen Thi Minh Huyen précise notamment le rôle du projet MAF :

« Le projet MAF (*Morphosyntactic Annotation Framework*) de l'ISOTC37/SC4 a pour but de définir un modèle générique dédié à l'annotation morphosyntaxique (norme ISO-24611). Ce modèle combine, d'une part deux niveaux de segmentation et de catégorisation linguistique (étiquetage morphosyntaxique), et d'autre part, un ensemble de catégories de données linguistiques permettant l'échange et l'interaction de données. Selon ce principe, les informations linguistiques (comme

---

[17] Le groupe de travail LISA/OSCAR a proposé des standards concernant, par exemple, l'échange de données de mémoire de traduction (TMX – *Translation Memory eXchange*), l'échange de données terminologiques (TBX – *TermBase eXchange*).

[18] Le site officiel de ce sous-comité se trouve au http://www.tc37sc4.org. Les documents de travail sont mis sur le site au fur et à mesure des activités du comité

les parties du discours, les traits morphologiques, *etc.*) de chaque annotation conforme au MAF doivent pouvoir être mises en correspondance avec les catégories de données définies. »

5. Les caractéristiques des documents en langue amazighe et/ou traitant de la culture ou des langues berbères :

Il s'agit notamment de ce qui est défini dans les axes de ce 4ème workshop de l'IRCAM. Évidemment, les axes proposés ne font pas le catalogue exhaustif des spécificités de la numérisation linguistique berbère (on ne peut sans cesse revenir sur les axes proposés dans les précédents workshops, ni même reprendre à l'identique les thématiques d'autres acteurs de la recherche berbère comme le CNPLET). Je commencerai par résumer de mon point de vue, les caractéristiques de ces ressources de documents ainsi que les objets de recherche (voire de pratiques[19]) qui me semblent correspondre aux objectifs de traitement correspondant aux différents usages qui peuvent être faits de ces ressources.

Les corpus berbères se caractérisent par leur aspect multimédia. On ne peut ignorer qu'une culture largement de tradition orale existe aujourd'hui, non seulement à travers des textes produits par des lettrés berbères, mais aussi des transcriptions réalisées par des ethnologues ou des linguistes. De nos jours, notamment grâce aux efforts d'alphabétisation en langue berbère, ces ressources linguistiques s'augmentent chaque année considérablement de par la production des journalistes, des pédagogues, des professionnels de la télévision ou de la radio, des chanteurs aussi, qui par leur production discographique constituent une partie considérable du corpus linguistique berbère moderne.

On l'a souvent souligné, les ressources linguistiques berbères existent sous trois traditions d'écriture : les tifinagh largement popularisés notamment au Maroc par l'effort spécifique de l'IRCAM, l'écriture arabe aménagée avec quelques caractères supplémentaires et l'écriture latine, elle aussi aménagée avec quelques signes diacritiques et d'accentuation. Grâce à la normalisation réalisée par l'ISO et

[19] On ne doit pas négliger que des communautés de producteurs ou d'utilisateurs de documents ne sont pas obligatoirement des chercheurs (ce peut être des cinéastes, des chanteurs, des enseignants de n'importe quelle discipline). Cependant, on doit admettre qu'ils interagissent tout autant que les chercheurs sur le (ou les) système(s) d'information lié(s) aux ressources berbères.

Unicode, sous l'initiative de l'IRCAM, cette diversité d'écriture a cessé d'être un handficap. Globalement, il est possible d'un clic de souris de passer de l'une à l'autre de ces formes d'écriture.

Il est par contre plus complexe de prendre en compte la diversité des différentes langues et différents parlers berbères. Couvrant un imense territoire sur une très longue période historique, les différentes ressources linguistiques berbères appartiennent à une famille de langues quelquefois non intercompréhensibles entre elles.

Cette diversité des types de ressources et de leurs versus de langue ou d'écriture, ainsi que la diversité des genres et types de médias déjà propres à tout corpus linguistique, imposent une rigueur de référencement lors de la mise en ressources numériques.

6. La nécessité de créer un header TEIberbère

Le Header TEI se divise en quatre sous-ensembles:

1. <fileDesc> *file description* : description bibliographique du fichier électronique XML-TEI (données utiles à l'indexation et au catalogage).
2. <encodingDesc> *encoding description*: description du projet et des choix éditoriaux d'encodage de la source (normalisation, corrections, traitement des fins de ligne, interventions éditoriales, etc.)
3. <profileDesc> *text-profile description*: la description des aspects non bibliographiques du texte (circonstances de la composition de la source, langue, sujet).
4. <revisionDesc> *revision description*: l'historique des révisions du fichier électronique.

Du point de vue de l'organisation numérique des ressources berbères, il est fondamental de disposer d'un header universel à même de pouvoir définir pour chaque document (ou sous ensemble de document) l'ensemble des différentes caractéristiques. Si je préconise un header TEI, c'est parce que je pense que dès cette étape la maîtrise et le paramétrage des documents ne sont pas seulement des tâches strictement techniques, mais doivent être le résultat du travail concerté des différentes communautés de chercheurs et utilisateurs. Certes avec l'aide indispensable de professionnels de l'informatique, mais aussi en sachant définir leurs propres besoins, tous les utilisateurs concernés devront pouvoir donner les

grandes caractéristiques des documents. : De quel document ou média s'agit-il ? Quel est son genre[20] ? quel est sa langue ou son parler ? Son écriture ? Est-ce un document original ou une transcription ? qui est responsable de ce document (auteur, éditeur, traducteur, transcripteur, commentateur…) ? Quel est l'auteur, la date, l'éditeur du document source ? Celle du document numérique (voire de ses différentes versions) ?

Dès cet étage du système d'information, on se rend compte qu'il faut pouvoir recueillir les différents desiderata des chercheurs et utilisateurs des ressources berbères, en faire la liste exhaustive et pouvoir ainsi disposer d'un étiquetage général des documents permettant de savoir quels types de traitements pourront être invoqués en fonction du type et genre de document et du type d'usages ou de recherches dont il relèvera. A l'évidence une transcription pétrographique ne sera pas traitée de la même façon que la transcription de témoignages oraux, des ressources pédagogiques d'alphabétisation, la numérisation d'un disque produit par un chanteur professionnel actuel… La mise au point d'un header TEIberbère ne pourra être réalisée qu'à l'issue d'une concertation entre des chercheurs et l'équipe porteuse du projet.

7. La définition de différentes communautés d'utilisateurs-chercheurs et des sous ensembles de balises TEIberbère qui leur sont spécifiques

En définissant ses différentes thèmatiques, l'appel à communication du 4è workshop montre bien les différentes familles de traitement nécessaires pour le traitement automatique des langues. Pour les grandes langues scientifiques ou industrielles, ce n'est que grâce à une mobilisation considérable de recherche développement en ingénierie linguistique qu'elles peuvent disposer aujourd'hui d'un environnement TAL. Il est de ce fait indispensable de procéder à une mobilisation «réellement collaborative et en ligne », de l'intelligence d'ingénerie linguistique permettant de résoudre puis de développer l'environnement TAL spécifique aux différentes langues de la famille berbère. Le problème n'est donc pas seulement de faire les bons choix techniques et normatifs, mais de les mettre à disposition d'un collège élargi de chercheurs linguistes et ingénieurs du langage à

[20] Texte manuscrit ou édité, disque, émission radio ou TV, transcription orale, transcription pétrographique, édition d'un écrit savant traitant de la langue ou de la culture, partition et paroles de chant

même de créer un niveau d'environnement TAL berbère opérationnel. Cela nécessitera des arbitages permettant de définir des degrés d'urgence et de priorités débouchant obligatoirement sur des choix de développement préférentiels.

Les précédents workshops de l'IRCAM ou des rencontres au CNPLET s'étaient mobilisés sur des axes thématiques qui correspondent aussi à des modes de structuration, de traitement ou d'analyse des ressources qui correspondent à des modes de balisages définis ou potentiellement descriptibles dans les guidelines de la TEI et ses grandes applications traditionnelles :

- L'analyse structurale et sémantique des textes
- L'e-learning
- La terminologie et la lexicographie
- La standardisation des langues berbères
- La réalisation de bibliothèques numériques berbères (ou au minimum franco-arabo-berbère)

8. Créer un groupe TEIberbère

La décision de créer un tel groupe TEIberbère à l'occasion du 4e atelier International sur l'amazighe et les TICs aurait, selon moi, plusieurs effets très stratégiques.

- La TEI est sans doute un des groupes « d'intelligence collective numérique » les plus anciens[21], les plus pérennes, et à ce jour très productif. Fonder une communauté TEIberbère aurait un effet d'émulation interne des chercheurs très bénéfique, mais aurait de plus l'avantage d'exciter l'intérêt de nombreux chercheurs non berbérophones appartenant à d'autres collèges disciplinaires de la TEI : construction des terminologies et dictionnaires, pratiques d'analyses littéraires ou d'analyse de corpus oraux, analyse de corpus bilingues, description des traits linguistiques, etc…
- Fonder un groupe de TEIberbère aurait aussi comme avantage de lier en une démarche unique la création du « collège international des chercheurs en culture amazighe » et la mise en chantier d'un sous-ensemble de guidelines

[21] Le consortium TEI a été créé en 1987 par trois sociétés savantes : l'Association for computers and the Humanities, l'Association for computational Linguistics, l'Association for Literary and Linguistic Computing. À l'heure actuelle, elle est structurée sous la forme d'un « TEI Consortium » qui est une institution sans but lucratif

spécifiques à cette recherche. C'est en cela d'ailleurs que réside ce qui fait le succès de la communauté TEI : mettre en synergie d'une part de la communication sociale[22] et d'autre part le chantier des « Guidelines TEIberbère » c'est-à-dire un lieu de ressources partagées, normalisées, interopérables permettant de formaliser sous forme numérique des méthodes et des savoir-faire indispensables pour la construction et l'exploitation de ressources langagières en berbère et à propos des langues berbères.

Sur chacun de ces deux derniers points, il y a un chantier de socialisation (tant présentiel que distant), un chantier linguistique et un chantier informatique à ouvrir. Il y a aussi (et surtout) un vaste chantier de formalisation et d'ajustement normatif des méthodes pour qu'elles puissent être repérées, puis formalisées dans des Guidelines TEIberbère :

a) Le chantier de socialisation est avant tout un travail de mise en collège relationnel (mais aussi hiérarchique ou par spécialité de métiers ou disciplines scientifiques) des participants de « TEIberbère ». Il existe un assez grand nombre d'outils à même de formaliser le fonctionnement participatif cependant il est nécessaire d'animer la mise en place du groupe (au début en utilisant un forum de discussion dédié) puis en paramétrant un outil pour rendre visible les groupes et sous groupes, notifier les hiérarchies de participation scientifique, autoriser la création de nouvelles balises, échanger des corpus, etc… La TEI dispose d'une grande expérience en la matière et le groupe TEI francophone (auquel je suis associé) pourrait très judicieusement accueillir TEIberbère.

b) Le chantier des Guidelines TEIberbère est un travail utile et urgent mais délicat et de longue haleine. Il ne s'agira pas de proposer la création anarchique et trop prolifique de balises spécifiques à notre sujet. Il faudra dans chacun des champs d'interventions proposés (analyse littéraire ou de corpus oraux, corpus multilingues, corpus terminologiques, description des traits linguistiques, et d'autres champs à définir) contrôler la réelle nécessité de création de nouvelles balises, mettre en place des sous-comités à même de proposer des procédures pour créer des consensus, éviter les doublons, proposer la fédération d'une famille de balises sous une seule balise déjà existante ou à créer avec des attributs multiples correspondant aux différents besoins spécifiques des chercheurs, etc…

---

[22] En paramétrant des mécanismes numériques formalisant le mode de fonctionnement du groupe humain des chercheurs (mais aussi des utilisateurs de ces recherches)

9. Quelle serait l'utilité de la TEI dans un tel contexte ?

Nous insistons sur la TEI *(Text Encoding Initiative)* car sa diffusion dans le milieu spécialisé des « humanitées numériques » (Digital Humanities) est en phase d'expansion. Un établisement comme l'Ecole des Chartes à Paris en fait maintenant un des axes primordiaux de l'enseignement du corps des archivistes français (qui fournit les hauts cadres des musées, des bibliothèques et des archives. Le monde anglo-saxon et germanophone s'est depuis assez longtemps approprié ce standard de traitement des ressources numériques.

On a vu aussi que c'était devenu le cadre de balisage normatif de la recherche développement en TAL.

Dans tous les domaines de la recherche linguistique et littéraire de nombreux auteurs ont déjà insisté sur la diversité des publics que doit viser une politique de création de ressources numériques : transformer une bibliothèque traditionnelle en bibliothèque numérique virtuelle implique des investissements et des dépenses de fonctionnement qui doivent être compensés par une valeur ajoutée des usages patrimoniaux ; en d'autres termes il est indispensable d'imaginer, puis d'aménager des nouvelles à facettes d'usages multi-spécialisés ciblées sur des publics nouveaux, internationaux, multilingues et multidisciplinaires.. Encore faut-il qu'une typologie de ces usages soit soigneusement repérée, étudiée, techniquement développée, puis culturellement et économiquement argumentée. Ce redéploiement théorique et social des usages nécessite bien sûr une ouverture des esprits chez les bibliothécaires et les documentalistes impliqués dans la gestion ou l'étude scientifique des ressources berbères, qui sont quelquefois très enracinés dans une seule des facettes d'usage, notamment celle très légitime d'un usage prioritairement ciblé sur des usages pédagogiques, et des usages qui faciliteront l'accès à un public arabophone ou francophone alors qu'un public très international pourrait dans certains cas être très interressé par des ressources berbères.

Cependant, la numérisation et la mise sur réseaux se rentabilise d'autant mieux que les usages ont été adaptés à la mondialisation réelle (multilinguisme, prise en compte du public international de la recherche berbère, attention soigneuse aux questions d'interopérabilité et de normalisation notamment pour les jeux de métadonnées attachées aux ressources, normes de codages des écritures). Ces questions de mondialisation des ressources numériques berbères sont souvent très techniques et heurtent bien légitimement les aspirations naturelles des « militants de la communication berbère » qui comprennent mal pourquoi ils devraient

consacrer de l'énergie à ces enjeux qui leur semblent un détour inutile pour parvenir à leurs objectifs propres.

En revanche, ces mêmes militants comprennent bien l'utilité évidente de disposer, sur le Net, en français (voire en anglais), de ressources documentaires leur permettant de repérer les recherches berbères entreprises partout dans le monde et pas seulement dans le périmètre naturel arabophone, francophone et hispanophone. En revanche, ils n'imaginent pas toujours que cet univers documentaire est, par construction, collaboratif et donc qu'il implique obligatoirement des réciprocités. La richesse du *Web* dans une discipline particulière n'existe qu'autant que les spécialistes concernés alimentent eux-mêmes le réseau avec leurs données propres en direction des autres communautés linguistiques. Cet échange inter communautés linguistiques s'applique aussi à d'autres facettes de l'échange : par exemple en direction d'autres pôles de synergie disciplinaire ou métiers (des musées[23],des médiathèques[24], des institutions dédiées à la pédagogie[25], à la littérature de jeunesse[26], des centres de recherche en ethnologie, en musicologie, etc.…

---

[23] Par exemple au Musée de l'Homme et d'Histoire naturelle (Museo de la Naturaleza y el Hombre) de Santa Cruz de Tenerife il existe des Momies Guanches. Les ressources d'une Bibliothèque numérique berbère auraient toute légitimité à être interrogeable soit par des chercheurs spécialisés, mais pourquoi pas par des visiteurs curieux d'en savoir plus sur cette civilisation berbère.

[24] La chanson berbère et la télévision berbère constituent une masse importante de documents qui existent déjà de toute façon dans de nombreuses médiathèques grand public, voire des archives spécialisées (phonothèques, cinémathèques) qui dans certains cas (Inathèque de France) ont déjà entrepris une numérisation systématique de tout ce qui est diffusé sur le territoire au titre du Dépôt légal. L'échange entre les chercheurs berbérophones et les conservateurs du Dépôt légal est ici évident : donner une grande valeur ajoutée à des ressources qui ne sont que balisées au niveau catalographique et disposer d'une numérisation systématique hors de portée des études berbères. Voir sur ce point particulier la contribution de Sabine Loupien, (Loupien S. 2010) .

[25] Dans le cadre du projet de Bibliothèque Numérique Franco-berbère soutenu par l'Organisation Internationale de la francophonie (Fonds Francophone des Inforoutes) des ateliers très spécifiques de pédagogie traditionnelle par les femmes en territoire touareg sont programmés et ils donneront bien sûr des ressources numériques pédagogiques. Dans le même projet ainsi que le projet de Bibliothèque Numérique Berbère (Ouhami Ould Braham et l'Alliance Cartago, dont le projet est soutenu par la Région Ile de France), seront

Cette situation n'est pas exceptionnelle et elle est bien connue des spécialistes des réseaux : le propre des réseaux est que la mutualisation mondiale (ou même nationale ou européenne) des ressources ne peut pas fonctionner de façon unilatérale.

Cette question de la mise en dynamique internationale et interdisciplinaire des ressources numériques est une question très universelle de la mondialisation numérique. La réussite de ces projets n'est évidente pour aucun secteur à l'exception de quelques domaines à haute valeur sécuritaire ou économique, ce qui n'est pas vraiment le cas du patrimoine berbère.

10. Créer une TEIberbère serait avantageux pour la communauté de la recherche berbère

Cela n'a rien d'une utopie. La communauté que constitue « l'initiative TEI » a totalement intégré XML dans des outils conformes aux « TEI guidelines » et on voit émerger de nouvelles générations d'outils permettant le paramétrage d'un modèle (c'est-à-dire. la mise en place d'un ensemble structuré de balises) dans un contexte XML. La communauté TEI, propose ainsi *« TEI Pizza Chief »*, un outil disponible en ligne et qui permet précisément de créer des DTD TEI adaptés à l'étude de tel ou tel corpus d'étude.

La constitution d'une DTD[27] est évidemment une opération exigeant un minimum de culture générale informatique (ou plutôt multimédia ; pas plus, en fait, que la

---

également prévues des Annales numériques du Baccalauréat Berbère en France (Animation de la recherche pédagogique Hocine Sadi.)

[26] Dans le cadre du projet de Bibliothèque Numérique Franco-berbère sera développé un volet important sur l'étude des contes berbères. Ces recherches animées par Janine Despinette et Tassadit Yacine seront directement relayées par le CIELJ (Centre International d'Etude de la Littérature de Jeunesse) et Ricochet (serveur spécialisé en littérature de jeunesse) à l'ISJM (Institut suisse jeunesse et média).

[27] *La Document Type Definition (DTD)*, ou Définition de Type de Document, est un document permettant de décrire un modèle de document SGML ou XML. Une DTD est la définition d'une SLG (structure logique générique). Puisqu'elle définit une SLG, une DTD détermine directement les contenus possibles pour un type de documents. Elle doit donc être élaborée avec le plus grand soin, et il existe plusieurs méthodologies, plus ou moins formelles, pour concevoir une DTD pour un type de documents donné.

maîtrise de la fabrication d'un site Internet). Cependant, il est important de bien considérer qu'une telle démarche de création d'un modèle TEIberbère ne saurait se réaliser sans la mobilisation des chercheurs en vue de construire avec eux les modèles de balises. Cette démarche d'analyse est strictement conceptuelle. Elle restera exclusivement centrée sur la définition des modalités fondamentales de la recherche linguistique et culturelle amazighe.

Ce travail de modélisation doit avoir pour objectif de concevoir un ensemble spécifique de balises organisées dans ce que les informaticiens appellent un schéma XML (ou DTD) et que le monde de la TEI qualifiera, une fois fait, de TEIberbère. Pour cela, il s'agit de recueillir auprès d'un ensemble représentatif de chercheurs les fonctions d'analyse savante, de pose de signets virtuels, de détermination de références, de pose de notes ou de gloses. Il ne s'agit pas de « martyriser les chercheurs traditionnels » pour les obliger à rentrer dans une démarche informatique, bien au contraire. Évidemment il serait opportun de disposer de quelques jeunes chercheurs, d'informaticiens et d'ingénieurs en industrie linguistique. La finalité de cette modélisation sera de comprendre, puis d'instrumentaliser en système numérique TEI, ce que font concrètement les chercheurs quand ils mettent en fiches, glosent et posent des signets dans des documents (un livre matériel réel), soit sur des œuvres ou des corpus plus globaux (musicologie, littérature orale, mass média berbères, ressources pédagogiques).

Ce travail de repérage des problématiques concrètes de recherche une fois réalisé, il s'agit dès lors de les formaliser, de les grouper et d'élaborer en consensus un modèle numérique qui constituera la nouvelle DTD de la TEI : TEIberbère.

Pour finaliser un tel projet, il est fondamental de modéliser les structures, les références et les zones que le chercheur veut qualifier au niveau sémantique. Il est aussi très important, en faisant ce travail d'analyse fonctionnelle, de s'assurer que ces nouveaux « projets de balises » concernent strictement la recherche berbère ne peuvent pas être récupérés (voire adaptés à l'aide d'attributs) dans d'autres domaines voisins de recherches (par exemple, comme on le soulignait, le TAL vietnamien, la normalisation des ressources linguistiques et de la terminologie, les études littéraires et les « humanités numériques ». Rappelons-nous, en effet, que par construction, le balisage de la TEI associe le noyau commun[28] à des jeux de balises additionnelles, si possible complémentaires les unes des autres.

---

[28] La TEI de base qui répond aux besoins des chercheurs en littérature ou de tout autre personne qui voudrait traiter des grands corpus de textes.

11. Pour amorcer un projet TEIberbère

Même si la communication proposée pour ce 4e Workshop n'est pas aussi courte et aussi peu technique qu'on l'aurait voulue à l'origine, sa présentation en séance sera aussi brève que possible (la moitié du temps qui me sera imparti).

Par contre un espace de questions et éventuellement de mobilisation des chercheurs dans une telle démarche pour créer TEIberbère devrait occuper les quelques 10 à 15 mn laissées libres du fait de la brièveté de l'exposé de ce projet.

Cette démarche est, me semble-t-il, particulièrement appropriée au genre « workshop » plutôt qu'à la démarche « colloque ».

Bibliographie

BURNARD L. "The Text Encoding Initiative: An Overview". Geoffrey Leech, Greg Myers, Jenny Thomas (eds.) *Spoken English on Computer: Transcription, Mark-up and Application*, 1995. London: Longman.

Digital Library Federation. *TEI Text Encoding in Libraries: Guidelines for Best Encoding Practices, Version 2.1 (March 27, 2006)*, 2007. <http://www.diglib.org/standards/tei.htm>.

IDE N., KILGARRIFF A., Romary L. "A Formal Model of Dictionary Structure and Content". *Proceedings of Euralex 2000*, (Euralex 2000) 2000. Stuttgart. pp. 113-126.

IDE N. & VERONIS J. *Présentation de la TEI*, in Cahiers Gutenberg n° 24, Juin, INRIA Rennes, 1996

LOUPIEN S. *Conserver et diffuser le patrimoine sonore kabyle de l'immigration : l'apport des métadonnées METS*, in colloque sur la Chanson kabyle, Paris, Cité de l'immigration, février 2010.

MYLONAS E., RENEAR A. "The Text Encoding Initiative at 10: Not Just an Interchange Format Anymore – But a New Research Community". *Computers and the Humanities* 1999. 33 (1-2) pp. 1-9. <http://dx.doi.org/10.1023/A:1001832310939>.

NGUYEN (Thi Minh Huyen) *Outils et ressources linguistiques pour l'alignement des textes multilingues français- vietnamien*, Thèse de l'Université de Nancy I sous la direction de Laurent Romary, 2006.

RENEAR A. « Theory and Metatheory in the Development of Text Encoding ». Michael A. R. Biggs, Claus Huitfeldt (eds.) *Philosophy and Electronic Publishing*, 1995.

<http://hhobel.phl.univie.ac.at/mii/pesp.html>.

VAUCELLE A. et HUDRISIER H. « Langages structurés et lien social », *tic&société* [En ligne], Vol. 4, n° 1 | 2010.

« http://ticetsociete.revues.org /790 »

Annexe 1 : A quoi ressemble un document TEI ?

Nota : La plupart des balises sont écrites en « écriture chameau » : le début de la balise est écrit en minuscule (c'est la tête moins haute que les bosses), puis tous les mots abrégés qui suivent sont écrits sans espace blanc avec une majuscule initiale.

```
<!DOCTYPE tei [ <!ENTITY TEI.prose "INCLUDE">]>
<tei>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Le plus petit document conforme à la TEI</title>
      </titleStmt>
      <publicationStmt>
        <p>Ce document n'est pas publié.</p>
      </publicationStmt>
      <sourceDesc>
        <p> Ce document est original.</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
 <body>
        <p>Voici le document conforme à la TEI le plus court qu'on puisse
imaginer.
      </p>
     </body>
  </text>
 </tei>
```

Annexe 2 : L'en-tête de la TEI (*TEI header*)

La description catalographique des documents numérisés est un aspect qui a été étudié en profondeur par un comité de la TEI. L'intérêt que suscite l'en-tête de la TEI d'un point de vue bibliothéconomique est certain. Tout en innovant, les solutions proposées s'harmonisent avec les processus déjà en place dans les bibliothèques.

L'en-tête de la TEI, qui fait partie de l'ensemble de balises obligatoires, sert à décrire un document balisé pour permettre aux utilisateurs d'avoir de l'information sur le texte lui-même: la (ou les) source(s), les principes utilisés pour le balisage et l'historique des révisions et modifications apportées au texte. Ces informations sont nécessaires autant pour les chercheurs qui utilisent les textes que pour les catalogueurs. Aucun document n'est conforme à la TEI s'il ne comporte pas la balise **<teiHeader>.**

Les 4 parties du TEI header

Les quatre parties de cet en-tête sont:

1- **<fileDesc>** peut être vu comme l'équivalent de la page titre d'un document papier. Il est difficile d'imaginer un document sans page titre, de la même façon l'élément **<fileDesc>** est le seul qui soit obligatoire pour la **<teiHeader>**. La flexibilité offerte par l'architecture de la TEI permet la description d'un texte en respectant la norme bibliothéconomique RCAA2.

2- L'élément **<encodingDesc>** décrit la relation entre le texte encodé et sa (ou ses) source(s). Il peut contenir, par exemple, de l'information sur le projet dans lequel s'inscrit l'encodage de ce texte ou des détails sur les décisions éditoriales qui ont été prises.

3- L'élément optionnel **<profileDesc>** permet de donner une description détaillée de ce qui caractérise les aspects non-bibliographiques du texte, telle la langue d'usage, la situation dans laquelle le texte a été produit, le nom des participants et leur rôle. La classification et les descripteurs assignés au texte font également partie de cet élément.

4- **<revisionDesc>** permet la description de l'historique des changements apportés au texte.

Annexe 3 : Quelques projets TEI dans le monde

Women Writers Project

Ce projet a débuté en 1989 à l'Université Brown. L'objectif est de constituer une base de données avec accès au plein texte de la littérature écrite par des femmes en anglais pour la période de 1330 à 1830.

***Center for Electronic Texts in the Humanities* (CETH)**

Mis sur pied conjointement par l'Université de Princeton et l'Université Rutgers en 1991, le CETH a pour objectif de promouvoir le développement, la diffusion et l'utilisation des textes électroniques en sciences humaines.

***The Oxford Text Archive* (OTA)**

Géré par les Oxford University Computing Services, l'OTA rend disponible plus de 1500 titres. Son site comprend des textes électroniques de plusieurs auteurs importants en grec, en latin, en anglais et en une douzaine d'autres langues.

American Verse Project

Il s'agit d'une nouvelle source de textes conformes à la TEI annoncée le 18 décembre 1995. Cette nouvelle initiative vient de *Humanities Text Initiative* de l'Université du Michigan et constituera une collection de textes de la poésie américaine.

•*Electronic Text Center - University of Virginia Library* (ETC)

Le ETC numérise et collecte depuis septembre 1992 des textes dans le but de les rendre disponibles par son service de textes en-ligne. Le Centre met également à la disposition de la communauté de l'Université de Virginie l'équipement informatique et les logiciels permettant l'analyse des textes tout en fournissant la formation nécessaire aux chercheurs pour l'utilisation de ces nouveaux outils.

Silfide Loria (Nancy)

Silfide (Serveur Interactif pour la Langue Française, son Identité,
sa Diffusion et son Etude); hébergé au LORIA Silfide recense tous les projets francophones utilisant la TEI[29]

---

[29] http://www.loria.fr/projets/Silfide/Index.html

Le projet DictAm
Dictionnaire électronique des verbes amazighe-français

Samira MOUKRIM

Laboratoire Ligérien de Linguistique-Université d'Orléans

samiramoukrim@yahoo.fr

1. Introduction

Dans le cadre de la promotion de la diversité linguistique dans la société de la connaissance, nous nous proposons d'élaborer un dictionnaire électronique des verbes amazighe-français (DictAm). Ce dictionnaire vise à rendre compte de l'ensemble des verbes dans le domaine berbère : *conjugaison, diathèse et sens*. Le DictAm a également une visée comparative dans la mesure où il rassemble et rend accessible les matériaux lexicaux des différentes variétés dialectales.

Le présent article a pour objectif de présenter le DictAm (**Dict**ionnaire Electronique **Am**azigh), un projet à travers lequel nous entendons produire une nouvelle ressource linguistique susceptible d'intégrer le berbère dans les nouvelles technologies de l'information.

2. Motivations

La langue berbère est parlée dans plusieurs pays d'Afrique (Maroc, Algérie, Tunisie, Lybie, Egypte, Mauritanie, Mali et Niger). Elle est partout minoritaire et diversifiée en de nombreuses variétés dialectales. Cette langue est aussi pratiquée au sein de l'Union Européenne (France, Allemagne, Pays Bas, Belgique, Italie, Espagne).

Pour diverses raisons socio-historiques et politiques, le berbère a connu un grand retard de la recherche linguistique. Le lexique reste le maillon faible des études berbères. Les outils lexicographiques disponibles semblent limités car dispersés et

les travaux existants sont partiels et ne concernent qu'un seul parler (ou dialecte). C'est la raison pour laquelle nous proposons un support *unique* à l'essentiel de l'information lexicale verbale berbère.

Comme beaucoup d'autres langues africaines, le berbère n'a guère bénéficié des avancées de l'informatique : un dictionnaire sous format électronique est très attendu. Par ailleurs, l'apprentissage du français par des berbérophones nécessite le développement d'outils didactiques qui prennent en considération leur langue maternelle. Le dictionnaire proposé peut être intégré également dans une perspective de didactique du berbère à des francophones.

3. Genèse du projet DictAm

L'idée du dictionnaire électronique des verbes berbères est née lorsque nous avons voulu évaluer le degré de variation au sein des verbes communs d'un certain nombre de parlers berbères[1]. Au début, nous avons utilisé des fiches (papier) pour classer les verbes collectés. Ce qui devenait de moins en moins pratique au fur et à mesure que le nombre de verbes et de parlers augmente.

Cela nous a poussé à réfléchir à la conception d'une base de données qui pourrait nous faciliter l'organisation de l'information et dans le même temps nous permettre de visualiser les convergences et divergences entre les parlers berbères étudiés.

4. Conception du DictAm

4.1. Méthodologie

Lors de l'élaboration du DictAm, nous nous sommes interrogée, d'une part, sur la démarche à suivre pour la structuration des données, et d'autre part, sur le traitement de la diversité linguistique.

En ce qui concerne la structuration des données, les verbes sont classés par ordre alphabétique de leur forme aoriste-impératif afin de faciliter la

---

[1] C'était en 2003-2004, après l'institutionnalisation de l'amazighe au Maroc et le début de sa standardisation.

consultation. Nous avons opté pour ce type de classement et non pour le classement par racine pour les raisons suivantes :

- Le DictAm s'adresse aussi bien aux usagers avertis qu'aux non-avertis (i.e. qui n'ont pas acquis les structures morphologiques élémentaires du berbère) ;
- Du point de vue de l'usage, il est plus aisé de chercher un mot en se référant à sa lettre initiale que d'en dégager la racine ;
- Le classement par racine présente un certain nombre de problèmes, en particulier lorsque celle-ci a subi des modifications au point de devenir méconnaissable[2].

La dimension bilingue du DictAm se manifeste au travers de l'association pour chaque entée lexicale berbère d'un équivalent en langue française. Par ailleurs, la structuration et le format des données ont été pensés de manière à permettre un transfert des données sélectionnées vers un document Word ou Excel (et prochainement HTML).

Quant au traitement de la diversité linguistique, le DictAm a été conçu de manière à *couvrir* le plus de parlers (et de dialectes) possibles et à *centraliser* toutes les données lexicales verbales des différentes variétés dialectales. La manière dont les données sont présentées permet de faire des rapprochements des différents parlers/dialectes (*cf. figures* 3, 5 et 6).

La structure de la base de données a été déterminée en prenant en compte toutes les caractéristiques du verbe[3] dans cette langue :

- En berbère, Le verbe peut être **simple** ou **dérivé** (causatif, passif, réciproque, etc.)
- Le verbe se présente sous trois thèmes principaux : l'**aoriste**, l'**inaccompli** et **l'accompli** (auxquelles nous pouvons rajouter l'accompli négatif et l'inaccompli négatif[4]).
- Les verbes sont généralement classés selon le nombre de consonnes radicales. On distingue plusieurs **types** de verbes : les verbes

[2] *Cf.* Taïfi (1990 : VI-XVI) pour plus de détails sur les différentes modifications que peut subir la racine en berbère.

[3] Pour déterminer la structure de la base de données, nous avons examiné les principaux travaux qui ont porté sur le verbe en berbère, ce qui nous a permis de prendre en compte toutes les caractéristiques du verbe.

[4] Et le prétérit intensif (résultatif) pour le touareg.

monolitères (constitués d'une seule consonne radicale), bilitères (2 consonnes), trilitères (3 consonnes), quadrilitères (4 consonnes) et quinquilitères (rares).

Toutes les informations concernant le verbe se présentent comme suit :

4.2. Matériaux

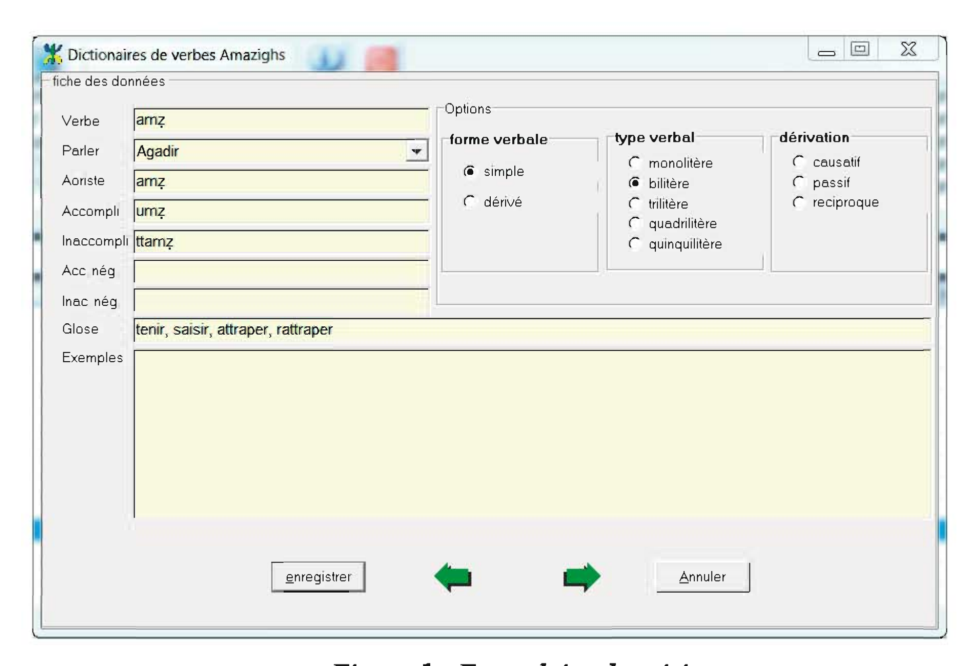


Figure 1 : Formulaire de saisie

L'alimentation de la base de données s'est faite à partir des sources documentaires existantes :

- Les dictionnaires classiques (version papier)
- Les lexiques et glossaires (accompagnant les descriptions grammaticales, recueil de textes, monographies…)
- Exploitation systématique des textes publiés

Actuellement, le DictAm comporte près de **3000 verbes** dans une

ctrl + d → ḍ
ctrl + s → ṣ
ctrl + h → ḥ
ctrl + g → g°
ctrl+Alt + t → ṯ
ctrl+Alt + d → ḏ

soixantaine de parlers berbères. C'est un travail qui est en cours de réalisation et qui a pour ambition de répertorier tous les verbes berbères ainsi que leurs équivalents en français.

Pour la programmation du DictAm, nous avons fait appel à El Amrani Mohammed, informaticien en Allemagne, qui nous a aidé à concrétiser ce projet :

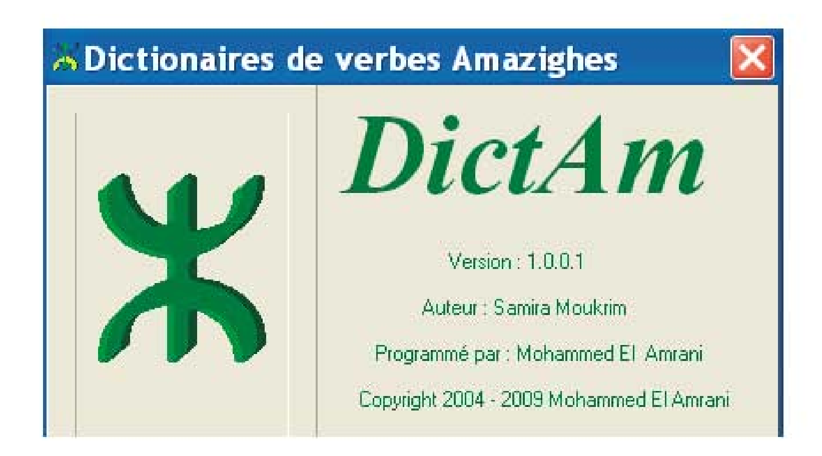


Figure 2 : Informations auteur et programmeur

4.3. Un nouveau clavier

Pour la notation du berbère, nous avons choisi dans un premier temps la graphie à base latine car elle permet une large diffusion.

Comme les claviers dont nous disposons sont faits pour les langues indo-européennes, nous avons crée un nouveau clavier (**DictAm_ARIAL**), avec *Microsoft Keyboard Layout Creator,* qui permet d'établir des raccourcis, pour taper directement au clavier les caractères spéciaux :

Travaillant dans une perspective de partage et de mutualisation, nous avons eu recours dernièrement aux polices Unicode, ce qui permet une large compatibilité avec les ordinateurs et logiciels récents.

Nous envisageons également d'introduire la graphie à base tifinaghe, en particulier après son intégration dans le standard Unicode (/ISO 10646).

5. Description du DictAm

Le principal intérêt du DictAm réside, d'une part, dans la rapidité d'accès aux données, et d'autre part, dans la possibilité qu'il offre de rapprocher des données issues de différentes variétés dialectales.

Toutes les informations concernant le verbe sont saisies dans le formulaire présenté dans la *figure1* ci-dessus. Dans l'interface de consultation, la fenêtre CHERCHER permet de répondre à n'importe quelle requête dès lors que cette interface permet de la formuler. Le fait de taper un verbe dans cette case permet d'accéder automatiquement à toutes les informations le concernant dans une multitude de parlers berbères, comme il apparaît dans la *figure 3*, qui présente le verbe **ddu** :

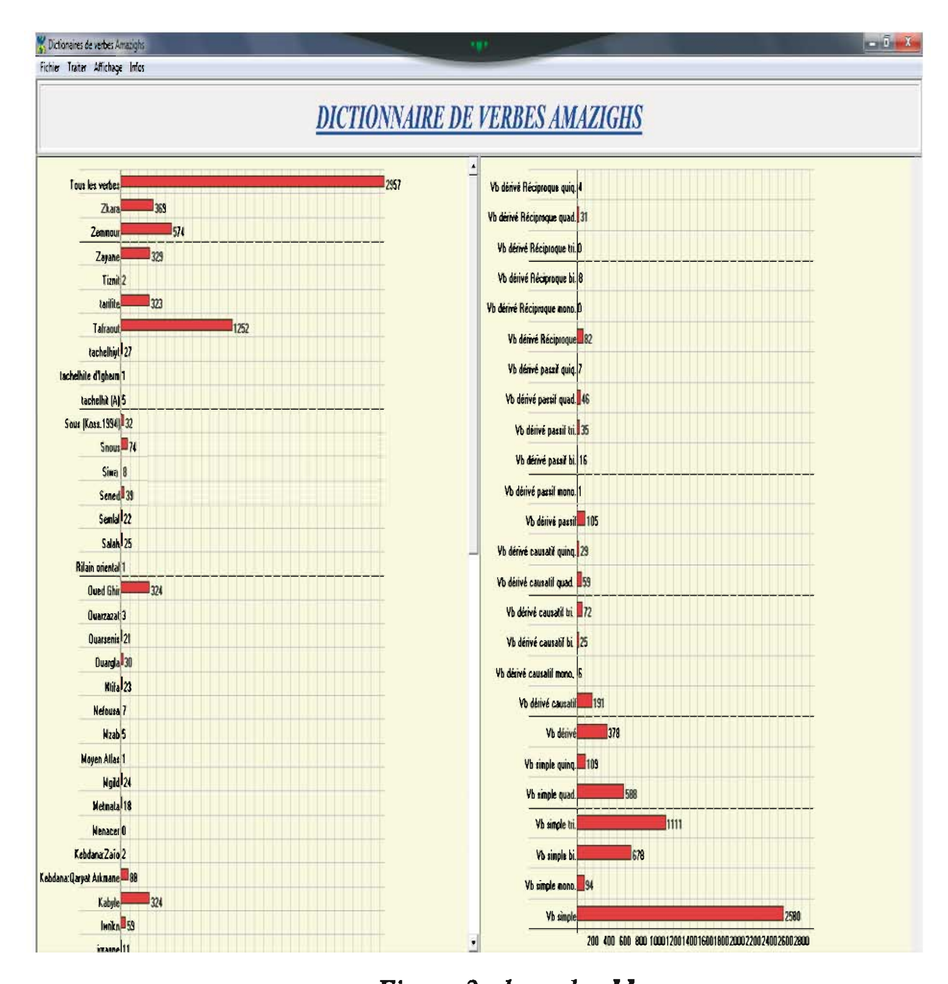


Figure 3 : le verbe ***ddu***

En ce qui concerne l'AFFICHAGE, nous pouvons faire un affichage normal (afficher tous les verbes dans tous les parlers) ou faire un tri :

- Afficher uniquement les verbes simples (tous les verbes simples ou un seul type de verbes simples : monolitères, bilitères ou trilitères, etc.)
- Afficher uniquement les verbes dérivés…
- Afficher les verbes communs à deux ou plusieurs parlers afin de les *comparer*, etc.

La fonction EXPORTER permet d'exporter les données, sous forme de texte ou de tableau afin de les imprimer ou les réexploiter via d'autres logiciels.

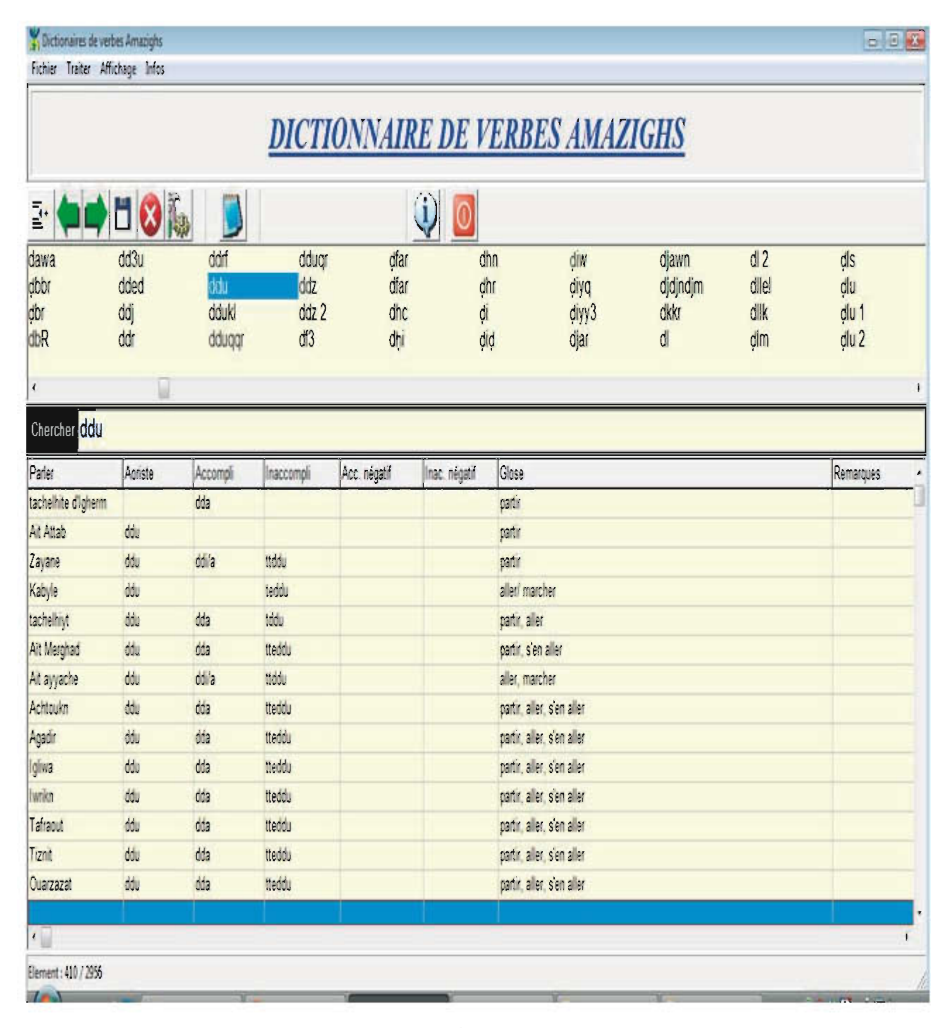


Figure 4 : Les statistiques

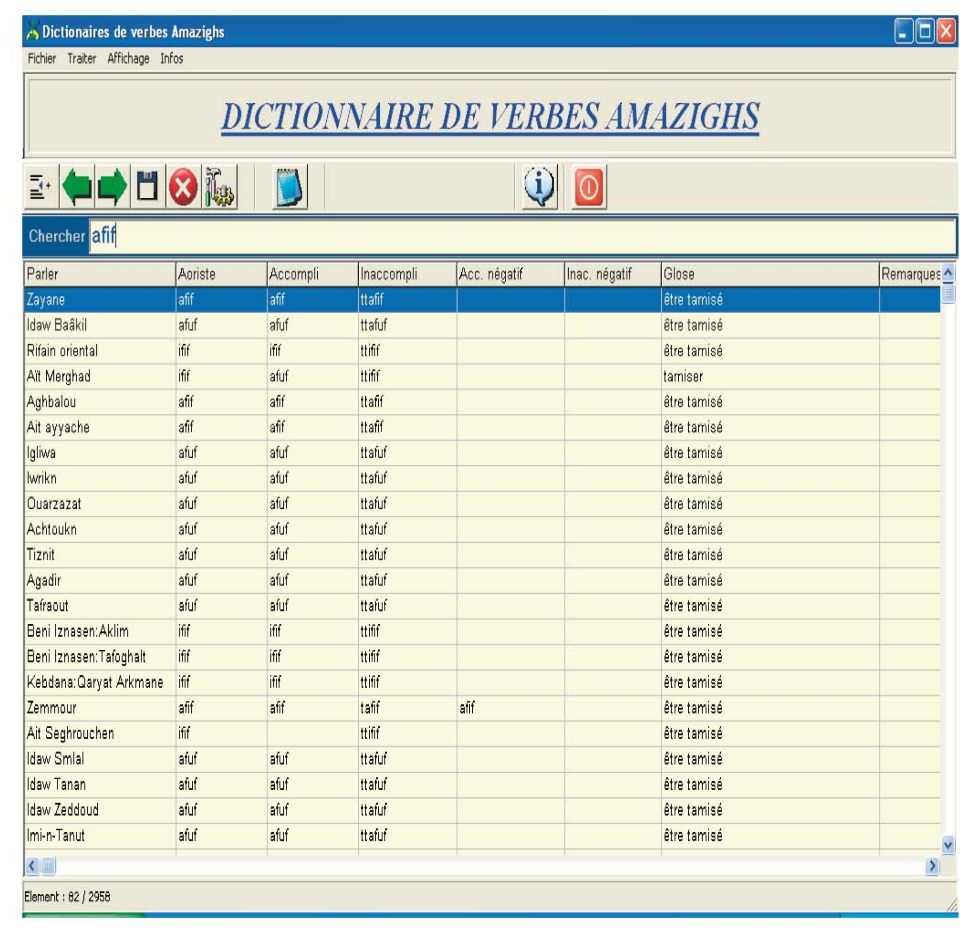
Nous avons ajouté récemment une nouvelle fonction : les STATISTIQUES[5], qui donnent un aperçu sur le nombre de verbes dans chaque parler, le nombre des verbes simples, le nombre des verbes dérivés, en fonction du type, etc.

---

[5] En construction.

6. Le DictAm : une visée comparative

Comme nous l'avons mentionné plus haut, la manière dont les données sont présentées permet de visualiser les convergences et les divergences entre les différentes variétés dialectales. A titre d'illustration, soit les verbes *afif* et *rar*, présentés respectivement dans les *figures 5 et 6* (infra). Ces deux verbes manifestent une variation de voyelle(s) et/ ou de schème(s) d'un parler à l'autre :



Parler	Aoriste	Accompli	Inaccompli	Acc. négatif	Inac. négatif	Glose	Remarques
Zayane	afif	afif	ttafif			être tamisé	
Idaw Baâkil	afuf	afuf	ttafuf			être tamisé	
Rifain oriental	ifif	ifif	ttifif			être tamisé	
Aït Merghad	ifif	afuf	ttifif			tamiser	
Aghbalou	afif	afif	ttafif			être tamisé	
Ait ayyache	afif	afif	ttafif			être tamisé	
Igliwa	afuf	afuf	ttafuf			être tamisé	
Iwrikn	afuf	afuf	ttafuf			être tamisé	
Ouarzazat	afuf	afuf	ttafuf			être tamisé	
Achtoukn	afuf	afuf	ttafuf			être tamisé	
Tiznit	afuf	afuf	ttafuf			être tamisé	
Agadir	afuf	afuf	ttafuf			être tamisé	
Tafraout	afuf	afuf	ttafuf			être tamisé	
Beni Iznasen:Aklim	ifif	ifif	ttifif			être tamisé	
Beni Iznasen:Tafoghalt	ifif	ifif	ttifif			être tamisé	
Kebdana:Qaryat Arkmane	ifif	ifif	ttifif			être tamisé	
Zemmour	afif	afif	tafif	afif		être tamisé	
Ait Seghrouchen	ifif		ttifif			être tamisé	
Idaw Smlal	afuf	afuf	ttafuf			être tamisé	
Idaw Tanan	afuf	afuf	ttafuf			être tamisé	
Idaw Zeddoud	afuf	afuf	ttafuf			être tamisé	
Imi-n-Tanut	afuf	afuf	ttafuf			être tamisé	

*Figure 5 : le verbe **afif***

Le verbe « **afif** » se présente à l'aoriste sous trois formes : **afif** (Aït Ayyache, Zayane, Zemmour, Aghbalou : Aït Messaoud), **ifif** (Aït Merghad, Aït Seghrouchen, Beni Iznassen d'aklim, Beni Iznassen de Tafoghalt, Kebdana de Qauriat Arkmane), **afuf** (Agadir , Iwrikn, Achtoukn, Tiznit, Ouarzazat, Igliwa, Tafraout). Dans tous ces parlers, il y a un syncrétisme entre l'aoriste et l'accompli à l'exception du parler de Aït Merghad ou l'accompli se réalise **afuf** au lieu de **ifif**. Quant à l'inaccompli, il est formé, dans tous ces parlers, par la préfixation de tt- à l'aoriste correspondant. Si l'on prend la forme de l'aoriste, par exemple, qui se présente sous trois formes selon les parlers :

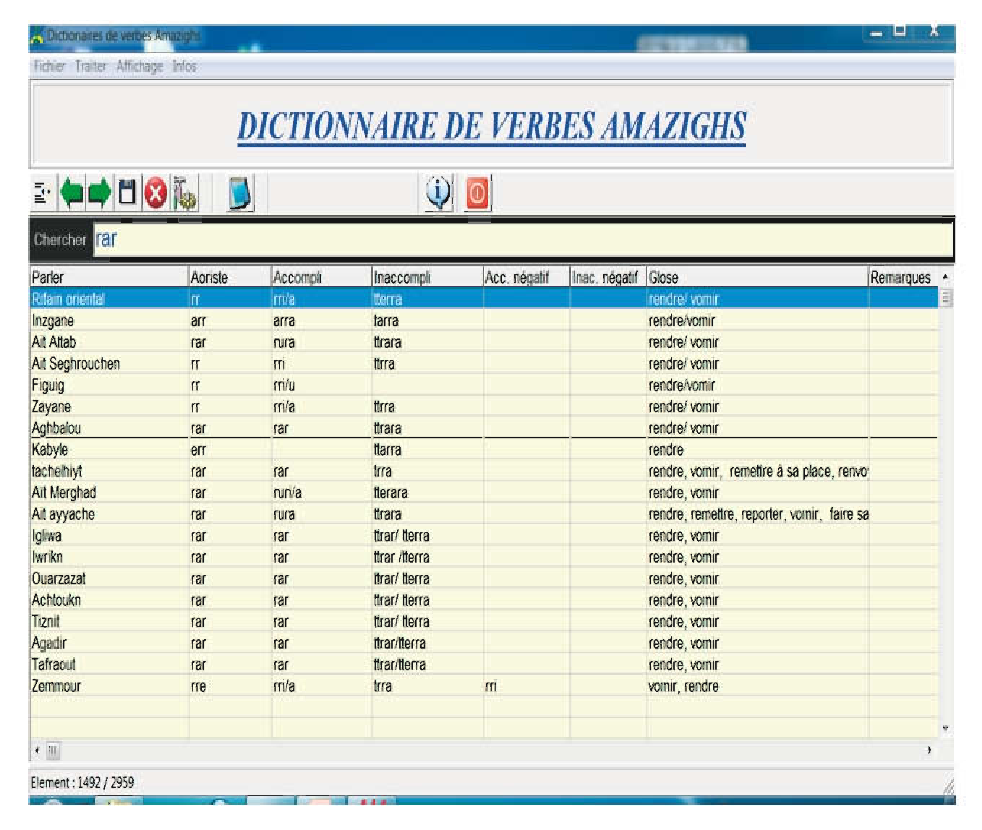
Aoriste	**schème**	**Voyelles**
afif	vcvc	a-i-
ifif	vcvc	i-i-
afuf	vcvc	a-u-

- Dans un processus de normalisation, quelle forme retenir ?

Deux solutions sont envisageables : soit on prend la forme d'origine i.e. la plus ancienne, soit on opte pour le critère de la représentativité dialectale, en choisissant la forme la plus usitée par le plus grand nombre de parlers. Bien qu'il ne soit pas toujours facile de trouver la forme d'origine de tous les verbes, nous pouvons opter pour la première solution avec la possibilité de recourir à la seconde dans le cas où la première s'avère impossible.

Si l'on examine les formes *ssiff*, *ttussiff* et *asiff*, avec lesquelles ce verbe est en rapport de dérivation, et qui sont respectivement sa forme factitive, sa forme passive et son nom d'action, nous constatons que dans la position pré-radicale –où la forme verbale **afif/ifif/afuf** présente soit la voyelle **i** soit la voyelle **a**– c'est la voyelle **i** qui semble être la plus ancienne. Quant à la voyelle intra-radicale –qui se présente sous deux formes **i** ou **u**– l'examen des formes *ssifif*, *afif* et *afifn* montre que c'est la voyelle **i** qui pourrait être la

plus ancienne. En effet, pour le verbe afif/ifif/afuf, les voyelles thématiques de base pourraient être i...i.. donc c'est la forme **ifif** qui pourrait être la plus ancienne. Les autres formes (afif/afuf) sont probablement le produit d'une évolution.



Dictionaires de verbes Amazighs

Fichier Traiter Affichage Infos

DICTIONNAIRE DE VERBES AMAZIGHS

Chercher rar

Parler	Aoriste	Accompli	Inaccompli	Acc. négatif	Inac. négatif	Glose	Remarques
Rifain oriental	rr	rri/a	tterra			rendre/ vomir	
Inzgane	arr	arra	tarra			rendre/vomir	
Ait Attab	rar	rura	ttrara			rendre/ vomir	
Ait Seghrouchen	rr	rri	ttrra			rendre/ vomir	
Figuig	rr	rri/u				rendre/vomir	
Zayane	rr	rri/a	ttrra			rendre/ vomir	
Aghbalou	rar	rar	ttrara			rendre/ vomir	
Kabyle	err		ttarra			rendre	
tachelhiyt	rar	rar	trra			rendre, vomir, remettre à sa place, renvo	
Ait Merghad	rar	run/a	tterara			rendre, vomir	
Ait ayyache	rar	rura	ttrara			rendre, remettre, reporter, vomir, faire sa	
Igliwa	rar	rar	ttrar/ tterra			rendre, vomir	
Iwrikn	rar	rar	ttrar /tterra			rendre, vomir	
Ouarzazat	rar	rar	ttrar/ tterra			rendre, vomir	
Achtoukn	rar	rar	ttrar/ tterra			rendre, vomir	
Tiznit	rar	rar	ttrar/ tterra			rendre, vomir	
Agadir	rar	rar	ttrar/tterra			rendre, vomir	
Tafraout	rar	rar	ttrar/tterra			rendre, vomir	
Zemmour	rre	rri/a	trra	rri		vomir, rendre	

Element : 1492 / 2959

*Figure 6 : le verbe **rar***

Si pour le verbe ifif (/afif/afuf), le schème sur lequel est construit le radical (vcvc) est le même dans tous les parlers examinés, ce n'est pas le cas du verbe rar/rr/arr (*cf. figure* 5). A l'aoriste ce verbe se présente sous trois formes correspondant à trois schèmes différents : **rr** (Aït Seghrouchen, Figuig, Zayane, Zemmour, Beni Iznassen d'aklim, Beni Iznassen de Tafoghalt, Kebdana de Qauriat Arkmane) / **rar** (Ayt Ayyache, Aghbalou : Aït Messaoud, Aït Merghad, Aït Attab, Agadir, Iwrikn, Achtoukn, Tiznit, Ouarzazat, Igliwa, Tafraout) / **arr** (Inzgane) :

Aoriste	schème	Voyelle
rr	cc	Ø
rar	cvc	-a-
arr	vcc	a-

A partir des formes dérivées de ce verbe : m**rara** , ssm**rara** , ttu**rar**, ta**rar**it, etc. nous pouvons constater que c'est la forme **rar** qui pourrait être la plus ancienne, de même que le schème correspondant (cvc).

Ainsi, le DictAm permet-il de visualiser et de rendre accessible les différentes formes sous lesquelles peut apparaître le verbe dans une grande partie de parlers berbères. Ce qui n'est pas sans importance pour la *standardisation* de la langue.

7. Mise à disposition et perspectives

Au terme du projet, la disponibilité et la diffusion des données auprès du public visé (chercheurs, étudiants, etc.) seront assurées au travers d'une interface Web déclinée dans les deux langues afin de renforcer l'accessibilité. Il sera possible également d'utiliser le DictAm dans les deux sens amazighe-français et français-amazighe. Une documentation simplifiée visant les internautes non avertis sera rédigée et mise en ligne afin de documenter la consultation des données.

Le DictAm peut être diffusé également au moyen d'autres supports, papier, CD-rom, clé USB, etc. Ainsi, les utilisateurs pourront le consulter en ligne ou hors ligne pour des usages aussi divers que l'éducation de base et l'enseignement en général, la traduction, la comparaison des variétés dialectales, et toutes autres activités en relation avec l'apprentissage ou la recherche.

Par ailleurs, il est prévu que des données audio viennent compléter le dispositif et, à terme, un fichier son, soit éventuellement associé à certaines entrées, quand cela est possible.

Enfin, nous envisageons, une fois le dictionnaire des verbes stabilisé, d'intégrer les autres unités du discours (les noms, prépositions, etc.) afin de construire un dictionnaire *général* du berbère sous format électronique.

8. Conclusion

Ainsi conçu, le DictAm répondra à trois types de besoins :

- Les besoins relatifs à la collecte et à l'organisation des données lexicales issues des différentes variétés de l'amazighe
- Les besoins des apprenants
- Les besoins des comparatistes et des chercheurs qui travaillent sur l'amazighe

En préservant toute la richesse héritée des différentes variétés dialectales et en intégrant l'amazighe dans les nouvelles technologies de l'information, le DictAm va sûrement contribuer à la *promotion* de cette langue.

Références

Azdoud, D. (1997), *Lexique commun des Ait Hadiddou du Haut Atlas*, El Jadida, Thèse de doctorat d'état. Faculté des Lettres et des Sciences humaines.

Boukous, A. (2003), « La standardisation de l'amazighe : quelques prémisses », *Standardisation de l'amazighe*, Actes du séminaire organisé par le Centre de l'Aménagement Linguistique à Rabat, 8-9 décembre 2003, Publication de l'Institut Royal de la Culture Amazighe, Série : Colloques et séminaires n°3.

Boumalk, A. et Bounfour, A. (2001), *Vocabulaire usuel de tachelhit, tachelhit-français*, Rabat, éd. Centre Tarik Ibn Zyad.

Cohen, M. (1969), *Essai comparatif, sur le vocabulaire et la phonétique du chamito-sémitique*. Paris, Ed. librairie Honoré Champion

Dallet J.-M. (1982), *Dictionnaire kabyle-français*, Paris, Selaf, (XI + 1056 p. ; 7ème vol. français-kabyle).

Delheure J. (1984), *Dictionnaire mozabite-français*, Paris, Selaf, (XVI + 322 p.)

Delheure J. (1987), *Dictionnaire ouargli-français*, Paris, Selaf

Dray M. (1998), *Dictionnaire français-berbère*. Dialecte des Ntifa, Paris, L'Harmattan, (510 p.)

El Mountassir, A. (2003), *Distionnaire des verbes Tachelhit-Français* (Parler berbère du sud du Maroc), Paris, L'Harmattan

Foucauld Ch. De (1951), *Dictionnaire touareg-français*, dialecte de l'Ahaggar, Imprimerie de France, 4 vol. (2028 p.)

Haddachi, A. (2000), *Dictionnaire de tamazight* : parler des Ayt Merghad (Ayt yaflman), Imp. Beni Snassen

Jordan A. (1934), *Dictionnaire berbère-français*, Rabat, Editions Omnia

Moukrim, S. (2003), *La flexion verbale en tamazight (parler de Zayane)*, Mémoire de DESA, Université Mohamed V, Rabat

Moukrim, S. (2009), « L'expression du présent actuel en arabe marocain, berbère tamazight et français, parlés à Orléans », in *Revue Sémantique et Pragmatique* (n° double 25-26)

Moukrim, S. (2010), *Morphosyntaxe et sémantique du « présent » : une étude contrastive à partir de corpus oraux, arabe marocain, berbère tamazight et français (ESLO/LCO)*, Thèse de doctorat, Université d'Orléans

Naït-Zerrad, K, (1999), *Dictionnaire des racines berbères*, Paris– Louvain : Edition Peeters.

Oussikoum, B. (1995), *Dictionnaire tamazight-français (parler des Ayt Wirra)*, Beni-Mellal, Thèse de doctorat d'état. Faculté des Lettres et des Sciences Humaines.

Rahho, R. (2005), *Dictionnaire berbère-français. Parler des Beni- Iznassen*, Fès, Thèse de doctorat. Faculté des Lettres et des Sciences Humaines, Dhar El Mehraz

Serhoual, M. (2002), *Lexicologie, lexicographie et sémantique berbères : 1) Lexicologie amazighe 2) Dictionnaire tarifit - français*, Tétouan, Thèse de doctorat d'état. Faculté des Lettres et des Sciences Humaines.

Taïfi M. (1991), *Dictionnaire tamazight-français* (Parlers du Maroc central), Paris, L'Harmattan-Awal, (XXII + 880).

Taine-Cheikh C. (2008) *Dictionnaire zénaga-français*, Berber Studies 20, Rüdiger Köpper Verlag. Köln

Projet de dictionnaire bilingue illustré (amazighe-français) des locutions nominales fauniques et florales

Malika CHAKIRI

Paris-Descartes-Sorbonne

malika.chakiri@parisdescartes.fr
chakirimalika@yahoo.fr

Résumé

Ce dictionnaire bilingue offre un florilège de locutions nominales en tamazight (parler du Moyen Atlas marocain) relatives à la faune et à la flore. Il est élaboré dans le souci d'éclaircir les difficultés des locutions figées et d'en guider l'usage. Nous entendons par locution toute suite polylexicale construite de monèmes non soudés formant un bloc figé inanalysable au niveau sémantique. Notre dictionnaire se veut clair et accessible pour tous. Il permet de comprendre un lexique spécifique que les locuteurs, et notamment les jeunes générations, ont tendance à oublier.

1. Introduction

L'objectif de cet article est de présenter quelques réflexions sur l'élaboration d'un dictionnaire spécialisé illustré bilingue amazighe-français. Notre démarche consiste à arrêter des repères méthodologiques dégagés lors de la constitution d'un glossaire assez riche de locutions nominales et verbales annexé à notre thèse de doctorat (Chakiri 2007).

L'article est organisé comme suit : Dans un premier temps, nous définissons les locutions et analysons leurs critères d'identification. Dans un second temps, nous présentons et analysons des exemples concrets relatifs à notre domaine d'exploration.

Pour la notation des données amazighes, nous utilisons le protocole suivant : - voyelles : *a*, *i*, *u* et *ə* pour noter le schwa. Semi-voyelle : *w*, *y*. Consonnes : *p*, *b*, *t*, *d*, *k*, *g*, *l*, *m*, *n*, *s*, *z*, *š*, *ž*, *ħ*/*ɛ* notent la fricatives pharyngales sourde et sonore, *x*/*ġ* les fricatives vélaires sourde et sonore, *h* la spirante, *q* l'occlusive dorso-uvulaire, *r* la vibrante apicale. Le point sous la lettre indique l'emphase, le *w* en exposant note la

labiovélarisation, le trait sous la lettre note la spirantisation, le dédoublement de la consonne indique la gémination.

Par ailleurs, les signes et abréviations adoptés sont : A. : aoriste, EA : Etat d'annexion, FM : formant, loc. : locution, N : nom, V : verbe, prép. : préposition, * : renvoie au défigement ou à des séquences non attestées dans la langue amazighe,

2 . Définition

Les locutions en tant que suite d'unités lexicales n'aboutissent à leur forme figée, intégrée dans le lexique et reconnue intuitivement et immédiatement, en synchronie, comme telles, par les locuteurs de cette langue, qu'en passant par plusieurs étapes. En effet, à l'origine, les locutions sont des créations individuelles. Elles se généralisent, lors des échanges verbaux et des interactions sociales, pour devenir ensuite, des expressions figées, formant ainsi une nouvelle unité dont le sens global diffère, le plus souvent, de celui de la séquence d'origine (Chakiri 2007, 2008). L'émergence de ces nouvelles unités composées et unifiées en un ensemble cohérent à partir d'unités lexicales ayant par ailleurs une existence autonome contribue à l'enrichissement du lexique tout en répondant au principe de l'économie linguistique.

3. Critères d'identification

Parmi les nombreux critères qui ont été soulignés par des grammairiens et lexicologues pour identifier les locutions, et que nous avons testé pour évaluer leur pertinence et leur caractère opératoire, nous avons retenu trois types de critères : morphologique, syntaxique et sémantique (Chakiri 2010).

3.1. Critères morphologiques

3.1.1. La polylexicalité

Ce critère renvoie à la présence d'une suite composée d'au moins deux monèmes ayant, par ailleurs, une existence autonome. Il constitue une condition nécessaire pour que l'on puisse parler de locutions : *poisson-chat*, *ddaw ṯaytt* « aisselle » (litt. « sous bras »). De ce fait, sont exclues, de notre champ d'investigation, les unités lexicales simples et les dérivées.

3.1.2. Anomalie lexicale

Ce critère est lié à la présence d'archaïsmes, c'est- à- dire d'unités désuètes non attestées dans le lexique en tant qu'unité autonome : *mamma ġyula* « cloporte » (litt. « maman l'ânesse »).

3.2. Critères syntaxiques

3.2.1. Non-prédication

La locution ne constitue pas une prédication :

- *un fait divers* -------**un fait qui est divers*

- *aġrum n tiġṯṯən* « champigon »-------**aġrum ddax n tiġṯṯən* (litt. « le pain, celui des chèvres »)

3.2.2. Blocage des propriétés transformationnelles

Ce critère concerne la possibilité ou l'impossibilité de manipuler les constituants de la locution :

- *Ṣabun n ṯəmġarin*

savon de vieilles

« Sortes de plantes savonneuses »

- **win ṯəmġarin ayd iyya Ṣabun*

ce savon est aux femmes âgées

3.2.3. Blocage des paradigmes synonymiques

Dans les syntagmes libres, on peut remplacer chaque unité lexicale par son synonyme. Cette liberté de substitution n'est souvent pas admise par les locutions :

- *Ṣabun n ṯəmġarin*

savon de vieilles

« Sortes de plantes savonneuses »

- **ṢṢabun n ṯuṯmin ṯimġarin*

savon de femmes vieilles

3.2.4. Non-insertion

Ce critère renvoie à la possibilité ou non d'insérer des éléments au sein de la locution :

- *illəs funas*

  langue bovin

  « buglosse »

- * *illəs axaṯar funas*

  langue grand bovin

3.2.5. Portée du figement

Ce critère permet de préciser les éléments sur lesquels porte le figement : il peut affecter la totalité de la locution. Dans ce cas, aucun constituant n'est libre comme dans :

bu nffax

celui qui a sifflet

« Cobra »,

c'est-à-dire que les deux constituants refusent toute modification ou transformation.

3.3. Critères sémantiques

Parmi les critères sémantiques spécifiant les expressions figées, nous avons retenu les quatre suivants :

- Unité de forme et de sens
- Compositionnalité vs non-compositionnalité
- Opacité vs transparence
- Motivation vs non-motivation

3.3.1. Unité de forme et de sens

Ce critère renvoie à l'unité sémantique de la locution : nous avons d'une part plusieurs signifiants et d'autre part un seul signifié. Cette dichotomie pluralité *vs* unicité a mené certains linguistes à conclure que les éléments constituants d'une locution disparaissent pour faire place à une image unique. Bien que ce critère soit devenu l'un des tests permettant de reconnaitre la locution, il demeure « trop difficile à constater, même par introspection, pour qu'on puisse le retenir pour

identifier ces complexes et les opposer aux syntagmes proprement dits » (MARTINET 1980). En effet, des locutions véhiculant deux ou plusieurs lectures ne répondent pas à ce critère. Elles peuvent évoquer une seule image chez les uns mais plus d'une image chez les autres, comme dans les exemples ci-dessous :

mm uġrum (sa)
1. « Poêlon en terre cuite servant à cuire le pain » (sé1)
2. « Vendeuse du pain » (sé2)
3. « Femme qui a du pain » (sé3)

awžžim n uġərḍay (sa)
1. « Plantain » (sé 2)
2. « La queue du rat » (sé1)

Certes, prises hors contexte, ce type de locutions admet plusieurs signifiés. Toutefois, en situation de communication, les locuteurs se réfèrent généralement au bon signifié, notamment s'ils partagent la même culture car, avant tout, ces locutions sont le produit de toute une société ayant ses propres normes et ses propres mœurs.

3.3.2. Compositionnalité et non-compositionnalité ; opacité et transparence

Une séquence est dite compositionnelle si son sens global est déduit de la somme du sens de ses constituants. Dès lors, toute séquence dont le décodage passe, sans poser de problèmes particuliers, par celui du sens de ses constituants est dite compositionnelle. A l'inverse, une séquence est dite non-compositionnelle lorsque le sens de chacun de ses formants n'intervient pas dans son sens global, en raison de l'absence de toute relation référentielle entre le signe linguistique et son référent.

Ainsi, dans les deux exemples suivants, où le premier est une locution française et le deuxième une locution amazighe (*pomme de terre*, *illəs funas* « buglosse »), bien que le sens de chaque constituant soit connu, la combinaison qui en résulte ne permet pas l'accès au sens de ces deux locutions ; elles sont non-compositionnelles. Cette non-compositionnalité est liée au phénomène de l'*opacité* car elle en est, en quelque sorte, le résultat. De ce point de vue, une locution dont le sens est déductible de la somme du sens de ses constituants, est dite transparente. En revanche, une locution dont le sens ne correspond pas à la concaténation du sens de ses formants est dite opaque.

3.3.3. Motivation vs arbitraire

Sur le plan du signe linguistique, l'arbitraire est défini par l'absence de toute relation référentielle entre le sens et la forme, entre le signifiant et le signifié. Mais, il n'en reste pas moins que dans une langue donnée, des signes par leur forme en rappellent d'autres.

Nous avons expliqué ci-dessus le rôle que jouent les éléments constituants dans la structuration du sens d'une locution. Or, pour que nous puissions décider si un élément d'une locution constitue le stimulus ou le déclencheur d'une telle dénomination, il faut d'abord connaître le sens de la locution. Une fois le sens détecté et le référent connu, nous décidons, à ce moment-là, du degré de motivation de la locution et de la contribution de ses éléments constituants dans son sens global. Si tous les constituants y contribuent nous dirons qu'elle est motivée, comme dans l'exemple ci-après :

ddaw ṯaytt
sous bras
« Aisselle »

Parallèlement, si aucun élément n'y participe, nous dirons que la locution est non-motivée ou arbitraire. L'esprit renonce dans ces cas à toute interprétation analytique. Les locutions de cette catégorie, *i-e*, les non-motivées ou les opaques nous imposent de penser l'arbitraire non comme un allié mais comme un ennemi. C'est le cas de *tête de mort*, *pomme de terre* ou encore,

aġrum n ṯġṭṭən
pain de chèvres
« Coprin micacé » (Champignon).

Cela étant, l'opacité n'est pas un critère suffisant car dans le cas des locutions transparentes, il n'est pas opératoire. En effet, bien que ces locutions soient transparentes, et probablement comprises d'un grand nombre de locuteurs, elles n'en appartiennent pas moins à la catégorie des expressions figées. En se basant sur ce genre de locutions, des linguistes s'intéressant au phénomène de figement relativisent l'idée que toute expression figée est opaque et le considèrent comme « un phénomène scalaire » allant de séquences totalement transparentes à des séquences totalement opaques, en passant par des séquences partiellement transparentes (Mejri 1997).

Ces tests effectués mettent en évidence l'impossibilité de varier tout type d'actualisation des éléments constituants des locutions, que ce soit le genre ou le nombre, ainsi que l'impossibilité de faire des transformations que peuvent subir les syntagmes libres. De ce fait, les locutions nominales expriment le plus haut degré de figement et relèvent du figement complet. Elles constituent, « un phénomène compact homogène ». De tels résultats ne font que confirmer quasi à coup sûr la conception absolue que l'on a des expressions figées et notamment les locutions nominales.

4. Typologie sémantique des locations

Partant des caractéristiques sémantiques des locutions citées ci-dessus et notamment du degré d'opacité sémantique, nous avons dégagé trois catégories de locutions.

4.1. Les locutions opaques.

La base sous-tendant la dénomination, dans cette catégorie, est absente car aucun élément de la locution ne fournit la dénomination. C'est pour cette raison qu'il est impossible de faire appel à la décomposition des éléments pour interpréter ce type de locutions. Elles sont considérées comme des séquences exocentriques car l'élément de base permettant l'accès à la dénomination n'est pas compris dans la locution :

illəs funas
élément de base = ∅
synthèse sémantique « buglosse »

La dénomination, dans ce type, est surtout fondée sur des motifs simples comme (Chakiri 2008) :

- La couleur du référent :

bu-ġmmu « rouge-gorge »,

renvoie à la couleur du cou de l'oiseau. *ġmu* « teindre ».

bu ħmran « rougeole »,

ħmər est un emprunt à l'arabe et signifie « rouge »

- forme et aspect

Ici, la figuration analogique est basée sur un simple rapprochement iconique avec le référent. Mais, comme nous l'avons mentionné ci-dessus, pour détecter ce rapprochement, il faut connaître le référent. Exemple :

aḍar n ufullus
pied de coq
« Pourpier »

- Activité

Dans d'autres locutions, la dénomination est basée sur le procès exprimé par le verbe qu'elles contiennent, en général un verbe d'action.

m nqqeb iẓẓuran « pic vert » (litt. Celui qui picore les racines)
m šərm igsan « mille pattes » (litt. celui qui ligote les chevaux)

- Effet et utilité

Dans cette catégorie, ont été regroupées les locutions qui soulignent l'effet ou l'usage domestique qui est fait du référent:

ṣṢabun n ṯəngarin,
(litt. « savon des vieilles (femmes) »)
« sorte de plantes savonneuse »,

autrefois utilisée par les femmes, la locution renvoie à l'usage qui est fait de l'objet.

- Propriétés médicinales

Dans ce cas, c'est l'effet curatif de la plante qui est retenu :

ħəbbətt rras
graine tête
« Dauphinelle », plante utilisée pour renforcer la pousse des cheveux.

- Moment de floraison

Des saisons et des moments de la journée servent de motif à la dénomination :

wərd lɛṢər
« Sorte de fleur qui s'ouvre en fin d'après-midi »,

lɛaṢər renvoie au moment de la prière qui a eu lieu aux alentours de 16h.

- Lieux d'existence

La dénomination détient sa motivation de l'endroit où le référent est localisé :

ṯulgəḏ aman
celle qui ingurgite eau
« Sorte de plante qui pousse au bord des rivières »

- Matière

hu ħbba
celui qui a une graine/ une balle
« Fusil »,

dans le cas présent « graine » est assimilée à « balle » par analogie de forme.

- Enfin, dans certains cas, des dénominations se fondent sur **des motifs complexes,** comme dans :

abrriḏ n ṯagdwin
bouc de pins
« Hulotte »,

deux motifs sous-tendent la création de cette locution : le lieu où vit l'oiseau et le cri de la hulotte qui rappelle le bêlement des boucs.

Toutefois, dans les locutions totalement opaques, certaines dénominations ont perdu toute relation référentielle avec leurs éléments constituants. Le référent ne fournit aucune information sur la motivation de dénomination. C'est le cas, par exemple, pour :

mamma ġyula « cloporte » (litt. « mère ânesse »)
ħəbb ləmluk « cerises » (litt. « grain des rois »)

4.2. Locutions totalement transparentes

Dans cette catégorie, la synthèse sémantique est le résultat de l'addition des sens des composants. Autrement dit, le sens locutionnel entretient un lien très étroit avec le sens de chaque élément constituant la locution. Par exemple, pour *ħəbb ššbab,* le sens global « acné » est déductible de *ħəbb* « grains ou boutons », et de *ššbab* « jeunesse » car « l'acné » se développe particulièrement pendant l'adolescence. D'autres locutions désignant des parties du corps :

ddaw ṯayt « aisselle »,
ṯiġmərṯ n ufus « coude »,

doivent leur dénomination à leur position dans le corps. Les locutions de cette catégorie sont dites endocentriques car « le signifié compositionnel est celui par lequel la locution fait sens » (Petit 1998).

4.3. Locutions partiellement opaques

Dites également « catégorie composite » (Chakiri 2007) car elle est constituée des conglomérés dont certains composants sont sémantiquement présents tandis que d'autres sont absents, et que la combinatoire des constituants peut fournir des indices permettant la lecture compositionnelle des locutions. Autrement dit, dans ce type de locutions, un élément demeure intact et conserve la base référentielle. Soient ces exemples que nous avons empruntés à Benveniste : *oiseau-mouche*, *chien-loup*, *poisson-chat*. Ici seul le premier constituant fournit la dénomination, la base sémantique demeure intacte car un *oiseau-mouche* est un oiseau, un *chien-loup* est un chien, un *poisson-chat* est un poisson. Dans ce cas, c'est l'expansion qui vient perturber l'analycité de la locution.
Dans cette catégorie, la détermination est, en général à gauche, c'est-à-dire qu'elle est fournie par le premier élément. Ainsi dans,
ṯabaxxuṯ n unzar « coccinelle » (litt. « Insecte de pluie »), le premier terme servant en quelque sorte de pivot autour duquel se construit le sens de l'unité complexe. *ṯabaxxuṯ n unzar* est un insecte. Le deuxième élément ne fait qu'apporter quelques informations ou spécifications liées, en général, à une certaine analogie iconique ou à des petits motifs, comme nous l'avons expliqué dans la catégorie I. Ces dénominations sont « en apparence membre [s] de deux classes distinctes qui pourtant ne sont ni homogènes, ni symétriques, ni même voisines » (Benveniste 1976).

5. Dictionnaire des locutions : éléments de méthodologie

Après avoir présenté et analysé les critères permettant l'identification des locutions, nous présentons dans ce qui suit notre méthodologie. Notre objectif est de rendre compte du lexique de la faune et de la flore à travers un travail lexicographique sérieux et illustré. Certes, il s'agit d'un lexique spécialisé et difficilement accessible pour tous les amazighisants non-initiés, mais il n'en reste pas moins qu'il s'agit bel et bien d'un type de lexique amazighe intégré dans la langue à travers ses usages. Ces expressions lexicales qui relèvent du « non ordinaire » méritent une analyse qui fournit les outils de leur décryptage et de leur compréhension.

Nous considérons que ce dictionnaire est novateur dans le sens où nous n'avons pas seulement la signification des locutions, mais nous avons tenu à ce que chaque locution, dans la limite du possible, soit accompagnée d'une illustration sous forme d'une image iconique représentant le référent. Ces illustrations servent à montrer, à quel signifié renvoient les locutions dont l'usage est en décalage avec l'emploi dit « standard ».

Précisons que ce dictionnaire n'affiche pas la prétention de définir toutes les locutions et comme peut le laisser entendre le titre, mais de définir le lexique le moins utilisé et le moins connu de la jeune génération. Environ 300 locutions expliquées et illustrées avec pour objectif central de se familiariser avec ce type de langage, de le pratiquer et de le comprendre linguistiquement et culturellement.

Il nous semble également pertinent d'émettre quelques remarques sur la présentation et le traitement scientifique du lexique. En effet, le but d'un dictionnaire étant d'être fonctionnel et compréhensible en éclairant ses lecteurs, nous adoptons dans ce dictionnaire une démarche explicative visant essentiellement le côté pratique et utilitaire. Pour en faciliter la consultation et vu la nature de locutions traitées, nous les avons regroupées selon la classe syntaxique à laquelle elles appartiennent. Au sein de chaque classe syntaxique, les locutions sont classées par l'ordre alphabétique de leurs constituants. Six structures syntaxiques ont été dégagées :

1. N + N[1]

bu *yəɛžižən*

celui qui a os

« Flamant »

[1] Nous avons également classé dans cette catégorie les locutions composés de ***bu*** + N

2. V + V

bbəy *rul*

couper + A se sauver + A

« Moustique »

3. V + N

ħərq *sus*

brûler + A carie

« Réglisse »

4. *m* + V + N

m ft̲əl t̲arfa

fm. faire des boules + A bouse

« Bousier »

5. N + prép. + N

aḏ̣il n tfiġra

raisin de serpent (EA)

« Bryone »

6. N + adjectif

t̲ifiġra t̲aħyuṭṭ

serpent folle

« Couleuvre »

La transcription des exemples est conforme, au moins pour l'instant, à l'alphabet phonétique international. Pour la traduction des composants de chaque locution, nous nous sommes basée sur notre connaissance de la langue en tant que locutrice native, quand cela a été nécessaire, nous avons fait appel à des spécialistes amazighophones. Chaque locution est suivie de deux traductions (traduction juxtalinéaire) (1) ; traduction littéraire/équivalent en français (2). Chaque locution est illustrée par une petite image référentielle.
Exemple :

illəs *funas*
1. langue bovin
2. « buglosse »

Les termes archaïques qui n'apparaissent que dans les locutions et qui sont méconnus de nos informateurs et des dictionnaires de la langue amazighe consultés sont présentés dans les traductions (1) par des pointillés - - - -:

ħəbb ršad

1. grains - - - -
2. « grains de cresson », « cresson alénois »

6. Conclusion

Nous avons essayé de présenter quelques points sous tendant l'élaboration d'un dictionnaire proche de la réalité parce qu'il met en avant l'univers culturel des amazighes du Moyen Atlas marocain à travers des blocs figés ayant pour corollaire la densité symbolique et culturelle. Toute personne ne partageant pas ou ne maîtrisant pas ce code symbolique et culturel ne peut parvenir à comprendre les motivations premières ayant engendré une telle locution, ni faire de rapprochements entre le sens locutionnel et les sens des constituants pris isolément. Notre souci majeur à travers ce travail est de permettre aux usagers de cette langue de partager les mêmes implicites socioculturels sans qu'il y ait de perturbation de décodage.

Références

BENVENISTE É. (1966). *Problèmes de linguistique générale*, tome I et tome II Paris, Gallimard.

BONHOMME M. (2006). *Le discours métonymique*, Berne, Peter Lang.

CHAKIRI M. (2010), « La locution nominale en berbère », Bayreuth-Frankfurt-Leidener Kolloquium zur Berberologie, Leiden, Pays-Bas, pp. 21-40.

CHAKIRI M. (2010), « Analyse stylistique des locutions nominales en amazighe », *Revue ASINAG*, IRCAM, Maroc. (à paraitre).

CHAKIRI M. (2007). *Le figement linguistique en berbère : aspects morphologique, syntaxique et sémantique*, Paris-Descartes.

CURAT H. (1982). *La locution verbale en français moderne. Essai d'explication psycho systématique*, Québec, PUL.

FONTANIER P. (1968), *Les figures de discours*, Paris, Flammarion.

GROSS G. (1996). *Les expressions figées en français*, Paris. Ophrys.

GROSS G. (1999). « Élaboration d'un dictionnaire électronique », *Bulletin de la Société de Linguistique de Paris*, Tome XCIV, Fasc. 1, Peeters, pp. 113-

138.GROSS M. (1982). « Une classification des phrases figées du français », *Revue Québécoise de linguistique*, n° 11, vol 2, Presse de l'Université du Québec, pp. 151-185.

MEJRI S. (1997). *Le figement lexical : descriptions linguistiques et structuration sémantique*, Publications de la Faculté des Lettres de Manouba, Série linguistique, vol. X, Université des Lettres, des Arts et des Sciences Humaines, Tunis 1.

PETIT G. (1998). « Remarques sur la structuration sémiotique des locutions familières », *Le figement linguistique*, Tunis, RLM, pp. 145-163.

POTTIER. B. (1992). *Sémantique générale*, Paris, PUF.

REY A. et al. (1997). *Dictionnaire des expressions et locutions*, Paris, Le Robert.

RAT M. (1957). *Dictionnaire des locutions françaises*, Paris, Larousse..

TAIFI M. (1991). *Dictionnaire tamazight-français* (Parler du Maroc central), Paris, L'Harmattan-Awal.

THUN H. 1975, « Quelques relations systématiques entre groupement de mots figées », *Cahiers de lexicologie*, n°2, vol XXVII, pp. 52-71.

Compiling of a Berber-French Dictionary (Figuig dialect)

Mohamed Yeou

Department of English & LERIC, Université Chouaib Doukkali
m_yeou@yahoo.com

The paper describes the compilation of a bilingual dictionary Berber (Figuig)-French both in paper and in electronic version. The dictionary is root-based and refers to dialectal forms for comparison. The purpose of the dictionary is to contribute to the documentation of Figuig Berber in order to provide a linguistic resource for the Figuigui community and scholars interested in researching the Berber language.

1. Introduction

1.1. Figuig

The Berber variety documented in this dictionary is spoken in Figuig and belongs to the Zenati branch of the Berber language family. Figuig is a oasis situated in the South East edge of Morocco, around 1000 km from Casablanca, and 460 km from the Mediterranean coast. The number of permanent residents is today around 15 000. Precise estimates of the language speakers is difficult to calculate because much of the population have emigrated away to Europe and to major Moroccan cities. The language has recently been listed in the UNESCO Atlas of the world's endangered languages. There is no doubt that such listing is justifiable given the following facts:

- the social and cultural context of the language has undergone great changes, resulting in lexical attrition;
- the language is used mostly by the parental generation and up;
- the limited numbers of L1 resident speakers, who are constantly subject to emigration;

- many speakers have negative attitudes towards the language and think their children would be better served by speaking other languages.

The present project of compiling a Berber-French dictionary is partly motivated by the critical need for language resource materials on which language revitalization and language standardization depend.

1.2. Review of literature

There are four major studies on the Berber of Figuig, as well as a few shorter ones. The earliest source is a small glossary of 31 pages by Basset (1885). However, Basset's work should be read with much caution. No more major fieldwork was conducted on Figuig Berber until 1994 and 1995, when Marteen Kossman and Fouad Saa defended their theses, respectively. Saa (1995) studies some aspects of the phonology and verbal morphology of Figuig Berber based on the generative framework. The thesis appendix is very interesting as it lists the verbal paradigm of 1296 verbs, along with their derived forms. Kossman's thesis, which was published as a book in 1997, gives a general description of the grammar of Figuig and provides a 144-page Berber-French glossary in the annex. This grammar is an excellent descriptive analysis and the glossary is very helpful. The final two major original sources are Ben-Abbas (2003) and Sahli (2008). Ben-Abbas (2003) investigates the sociolinguistic aspects of word-borrowing between languages in contact, mainly Arabic and French, while Sahli (2008) gives a brief grammar sketch of the language, together with a Berber-Arabic glossary. The glossary consists of a list of 2250 words without context, exclusively from the dialect spoken in Ksar Laâbidate. Another important study is a collection of folktales by Ben-Amara (2007). Ben-Amara transcribes an interesting number of tales from Figuig, but does not translate them. He gives, however, a list of the words used in these tales with their French gloss.

2. Printed Version of the dictionary

2.1. Compiling of the dictionary database

The dictionary database is compiled using Toolbox, a program produced by SIL International (formerly the Summer Institute of Linguistics). Toolbox uses MDF standard (Coward and Grimes, 2000) for lexicon structure and converts certain files in Standard Format into RTF (to be further processed and printed with MS-Word). Toolbox provides field lexicographers with the ability to integrate various types of data: lexical, grammatical, semantic, etc. It has many options for selecting, sorting, and displaying data. It is very useful for helping researchers

generate a reversed finder list as well as analyze and interlinearize text corpora. A sample of the printed output for a formatted dictionary and a reversed finder list are found in the appendix.

The typical database entry has the following fields:

\lx	Lexeme (is the abstract consonantal root)
\se	Subentry (derived word)
\va	Variant form
\vn	Variant comment (shows source of variant: name of Ksar or name of author)
\ps	Part of speech
\sn	Sense number
\rn	Reversal (this gives the French word(s) or phrase(s) desired for a reversed French-Berber finder list)
\dn	French definition
\sc	Scientific name (two-part name of a species, especially for plants)
\ng	Grammatical information (mainly for the different verb stems)
\sy	Synonym
\an	Antonym
\cf	Cross-reference (general purpose cross-reference)
\xv	Example in Berber
\xn	Translated example in French
\sd	Semantic domain (for entering semantic fields)
\nt	General Notes (dialectal forms and language name from which the word is borrowed; dialectal forms come from published Berber dictionaries and glossaries)
\vr	See (this is a field which I added to cross-reference a variant item to a main entry where fuller information is found)

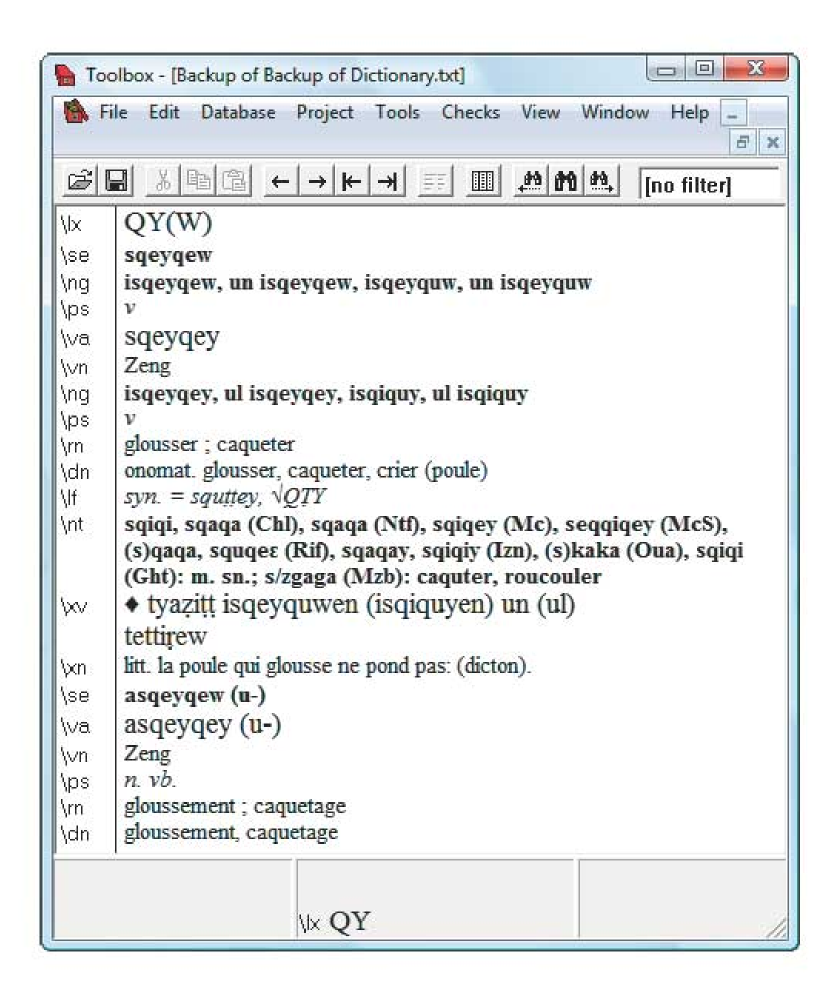


Figure 1: sample entry from the dictionary database (in Toolbox format)

A sample entry which shows how the information in database is organized (in Toolbox format) is given in Figure 1. The \lx field shows the main form of the lexeme, which is actually the consonantal root, and \ps gives the part of speech. \se shows the derived words. The \xv field gives the definition of the word in French. \xv, \xn pairs give example sentences and their French translations. \va lists variant transcription of the item, and \vn shows which source the variant came from (name of Ksar). The \nt field lists attested Amazigh dialectal forms.

2.2. Amazigh Lat Keyboard

The transcription of Berber in the dictionary is based on a standardized Latin transliteration system, as used by Berberists. However some of the symbols can be different from what is commonly used, as shown in (1):

(1)

Our system	Other systems	IPA
e	ə	ə
š	c	ʃ
ž	j	ʒ
tš	č	tʃ
dž	ǧ	tʒ
ɣ	ġ	ɣ
x	ḫ	χ
d̥	d / t	t
t	t / t^{s}	t^{h}

To enter text in Berber, I do not use Windows tools such as Character map because they can be very cumbersome. Instead, I created a keyboard layout designed to my own specification. The virtual keyboard, called Amazigh Lat, is made for regular writing and is compatible with Unicode fonts, such as the most recent versions of Times New Roman (from v. 5.0) and Arial (both supplied in Windows Vista and Windows 7). Amazigh Lat keyboard is used with Tavultesoft Keyman Desktop program, a utility for managing keyboard input methods. Both the program and the keyboard can be downloaded at http://www.tavultesoft.com.

2.3. Transcription approach

The transcription adopted for lexical words in the dictionary is phonetic (broad) rather than phonological. The intent is to show phonetic aspects that characterize Figuig variety and compare it to other Amazigh varieties. Apart from spirantization, devoicing of some geminates is an important feature that is noted at the phonetic level. The voiced geminates /bb, dd, ḍḍ, gg, gg^{w}/ are realized as [pp, d̥d̥, ṭṭ, kk, kk^{w}] both at the lexical and morpho-phonenic level. The voiceless non aspirated geminate [d̥d̥] (IPA [tt]) is indicated through the use of the IPA diacritic for devoicing, the under-ring, to distinguish it from the voiceless aspirated geminate [tt] (IPA [tt^{h}]) (Yeou et al., 2011).

The phonetic approach is also motivated by the intent to record dialectal variation inside Figuig Berber itself. For example, the data below shows that

absence / presence of pharyngealization or aspiration in an important indicator of dialectal variation:

(2)	*Upper Figuig*	*Zenaga*	
	ttrid	d̥d̥rid	"thin pancakes"
	d̥d̥ir	deyyer	"foot of mountain"
	iɣ zer	iɣ ẓeṛ	"river, wadi"
	tšar	.tšaṛ	"fill (up), be filled (up)"
	d̥d̥er	tter	"live"
	tter	tter	"ask for charity"

As far as the representation of morpheme boundaries is concerned, I adopt a 'syntactic' approach based on the criterion of syntactic mobility (see Stroomer, 1994). Hence spaces will preferably used rather than hyphens to mark morphological segmentation. For example, transcription type (a) will be adopted rather than type (b):

(3) 1a. Inna yas sad isek tid̥d̥art nnes "He told him that he will build his house"

1b. Inna-yas sad isek tid̥d̥art-nnes

2a. ppas d mmis ṛaḥen ɣ res "His father and his son went to see him/her"

2b. ppa-s d mmi-s ṛaḥen ɣ r-es

As regards assimilation processes, I generally adopt a phonological approach:

- The subjunctive and future marker *ad* is always transcribed as /ad/, given that its assimilation to the following consonant is predictable: *ad tɣ er* [at tɣ er] "that she will study", *ad nɣ er* [an nɣ er] "that we will study"
- The prefix /t-/, which assimilates in voicing to the following consonant, is also transcribed phonologically: *tezdeɣ* [dezdeɣ] "she lived", *težžey* [dežžey] "she recovered"

However, a phonetic approach was preferred in the case where there might be some dialectal variation, namely for the feminine morpheme suffix /-t/, the causative prefix /s(s)-/, and the intensive form prefix /tt-/:

(4)	*Loudaghir, Maïz*	*Other ksour*	
	tadist	tadiss	"stomach"
	tamaẓuẓt	tamaẓuṣṣ	"late season, youngest daughter"
	tašemmušt	tašemmušš	"bundle, knot"

(5) *Upper Figuig* *Zenaga*

	sḍuṛḍeṛ	ṣḍuṛḍeṛ	"deafen"
	ssenḍew	ṣṣenḍew	"cause to jump"
(6)	*Loudaghir*	*Zenaga*	
	ttezdid	ddezdid	"become thin [intensive form]"
	ttezluluf	ddezluluf	"singe [intensive form]"

2.4. The layout of the entries

The present dictionary uses a root-based approach, even if this approach has its practical weaknesses. This approach was partly motivated by the desire to serve the needs and interests of the academic community of linguists interested in researching the Amazigh language, or related languages.

The layout of the entries is organized as follows:

- Roots in bold capitals are arranged in the following alphabetic order. B, D, Ḍ, F, G, H, I, K, K^{w}, L, M, N, Ɣ, R, S, Ṣ, Š, T, Ṭ, W, X, Y, Z, Ž, Ɛ. Roots starting with Ḍ, Ṛ and Ẓ are, however, listed with D, R and Z, respectively
- The items in a root entry are grouped according to their semantic relation.
- For each root entry, simple verbs are listed first, and then the derived verbs with the following prefixes: /s-/, /m-/, /ttw-/, for the causative, the reciprocal and the passive, respectively. After that, verbal nouns for both simple and derived verbs are given, and finally, nouns and adjectives.
- For each verb, the first line gives the aorist as the basic form, followed by the other forms, mainly the preterite, the negative preterite, the intensive, and the negative intensive. The slash separates variant forms if there are any.
- For each noun or adjective we list the marker of the construct state between parentheses: (u-), (w-), etc.

2.5. The strengths of the present dictionary

2.5.1. Exhaustiveness and volume

This dictionary project aims to develop a comprehensive dictionary of Figuig Berber with French translation and extensive dialectal cross-references. The present dictionary tries to avoid the following weaknesses of previous work on the lexis of Figuig: (1) word-for-word translation; (2) lack of real uttered sentences; (3)

limitation to one community dialect; (4) limitation to the literal dimension of meaning

In this project, I will bring together not only an extensive compilation of words in Figuig with authentic sample sentences and their French translations, but also figurative and idiomatic uses of some of these words. In addition to that, some patterns of expression like proverbs, riddles and excerpts from songs and tales are included, because they reflect the culture of the Figuigui community more than every other kind of linguistic unit.

The Amazigh Figuig variety is characterized by some minor dialectal variation due to the fact that Figuig comprises seven seperate *ksours* or communities situated on two levels: The upper level consists of Laâbidate (At nnež), Loudaghir (At □ed̥d̥i), Oulad Slimane (At slimane), Hamam Tahtani (At wad̥d̥ay), Hamam Foukani (At □ameṛ) and El Maïz (At lem□iz), and the lower level consists of Zenaga (Iẓnayen). There is complete mutual intelligibility across the communities, and the small variation that exists will be noted in the dictionary. The default dialect is upper Figuig, namely Loudaghir, but variant forms are listed in the variant field (\va).

2.5.2. Dialectal cross references

At the bottom of each lexical entry, the dictionary lists attested forms from Amazigh languages or dialects given in Table (6), and whose references are given in the bibliography. It also indicates if the meaning is different or similar to that of Figuig. If the word is borrowed, etymological information about the source language and the original form is given.

(7)

Mc	Tamazight of central Morocco	Zen	Zenaga of Mauritania
McS	Tamazight of south central Morocco	Aha	Tahaggart
Chl	Tachelhit	Nig	Tamajeq (Tawellemmet, Tayert)
Rif	Tarifit	Mal	Tamasheq of Mali
Izn	Beni Iznassen	Nef	Nafusi
Sen	Senhaja de Srair	Ght	Ghat
Ntf	Ntifa	Ghad	Ghadamès

Kab	Kabyle	Snd	Sened
Che	Chenoua	Djr	Djerba
Cha	Tachawit or Chaoui	Chn	Chenini
Sns	Beni Snous	Dw	Douiret
Ace	Central Algeria	Ght	Ghat
Mzb	Tumzabt of Mzab	Ghad	Ghadamès
Oua	Tagargrent of Ouargla	Siw	Siwa
Tim	Gourara, Touat, Tidikelt		

3. Electronic Version of the dictionary

To generate the electronic version of the database, Lexique Pro was used. Lexique Pro is a free program developed by the Summer Institute of Linguistics (SIL). It transforms data from a Toolbox and formats it in a dynamic viewer.

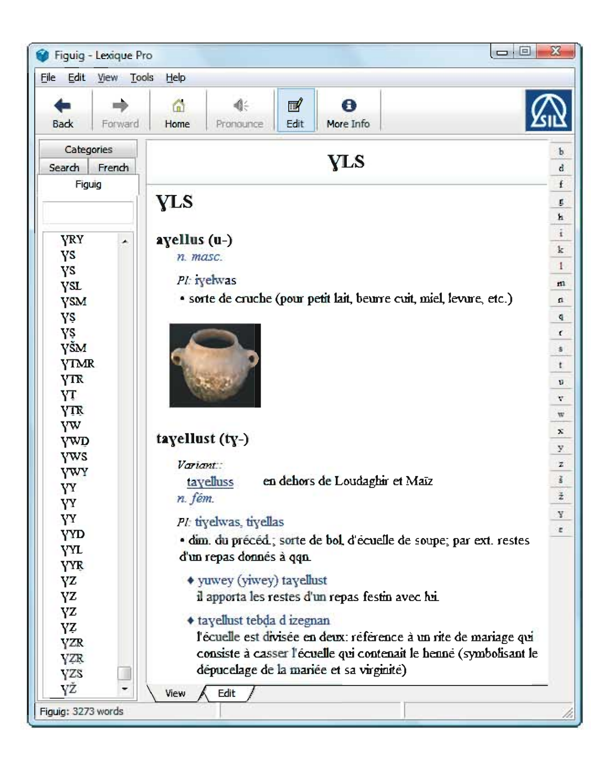


Figure 2: snapshot of the electronic version of the dictionary

The generated interactive dictionary has the advantage of displaying the database in a user-friendly format. The user can navigate by clicking on hyperlinks to related entries such as synonyms, variants, cross-references, semantic categories. Figure 2 shows a snapshot of a lexical entry from the electronic version of the dictionary. On the left we can either access the entries, the dictionary roots, by clicking on "Figuig", or access the reversed French finder list by clicking on "French". Both the roots and the French words are organized in alphabetical order. Moreover, we can also view entries by semantic domain if we click on the Category button. Work on thematic categorization is ongoing and only the following semantic fields have been entered so far :

- Kinship terms
- Animals
- Body Parts
- Food & Drink
- Clothing
- Agriculture & Vegetation
- Religion & Beliefs

The advantage of the electronic version generated by Lexique Pro is that it can be distributed as a packaged setup program and can also be exported to a set of web pages, in text, html, or xml formats.

4. Conclusion

The dictionary has been on the making for several years. We hope that it might be completed in approximately six months. It is expected that the dictionary will have roughly eight thousand words when completed.

5. Bibliography

Adnor, A. (2004). *An Electronic Tashlhit-English dictionary (prototye).* Thèse de Doctorat d'Etat. Rabat: Université Mohamed V.

Alojaly, G., Prasse, K.-G., Ghoubeïd A. and Ghabdouane M. (2003). *Dictionnaire touareg-français.* Copenhagen: Museum Tusculanum Press.

Amaniss, A. (non publié). *Dictionnaire tamazight-français.*

Azdoud, D. (2011). ***Dictionnaire berbère-français.*** **Paris:** Maison des sciences de l'homme.

Basset, R. (1885). Notes de lexicographie berbère: dialectes des k'çours Oranais et de Figuig. *Journal Asiatique,* 8/6, 302-71.

Beguinot, F. (1942). *Il Berbero Nefûsi di Fassâto. Grammatica, testi raccolti dalla viva voce, vocabolarietti.* Roma: Istituto per l'Oriente.

Ben Abbas, M. (2003). *Variation et emprunts lexicaux : étude socio-linguistique sur le parler amazigh de Figuig*. Thèse de Doctorat. Fès: Université Sidi Mohamed Ben Abdellah.

Ben Amara, H. (2006). *Tanfust : Recueil de récits amazighs de Figuig*. Rabat: Publication du Ministère de la Culture et l'IRCAM.

Bounfour, A. and Boumalek, A. (2001). *Vocabulaire usuel du tachelhit*. Rabat: Centre Tarik Ibn Zyad.

Cadi, K. (1987). *Système verbal rifain : forme et sens (Nord-Marocain)*. Paris: SELAF.

Coward, D. F. and Grimes, C. E. (2000). *Making dictionaries: a guide to lexicography and the Multi-Dictionary Formatter, version 1.0*. Waxhaw, North Carolina: Summer Institute of Linguistics.

Dallet, J.-M. (1982). *Dictionnaire kabyle-français*. Paris: SELAF.

Delheure, J. (1984). *Dictionnaire mozabite-français*. Paris: SELAF.

Delheure, J. (1987). *Dictionnaire ouargli-français*. Paris: SELAF.

Dray, M. (1998). *Dictionnaire français-berbère. Dialecte des Ntifa*. Paris: l'Harmattan.

Destaing, E. (1914). *Dictionnaire français-berbère*. Paris: l'Harmattan. 2nd edition.

Destaing, E. (1938). *Etude sur la tachelhit du Sous, vocabulaire français-berbère*. Paris: Leroux.

El Mountassir, A. (2003). *Dictionnaire des verbes Tachelhit-Français*. Paris: L'Harmattan.

Foucauld, C. de (1951-1952). *Dictionnaire touareg- français: dialecte de l'Ahhagar*. Paris: Imprimerie Nationale.

Gabsi, Z. (2003). *An outline of the Shilha (Berber) vernacular of Douiret (Southern Tunisia)*. PhD. Thesis. University of Western Sydney.

Heath, J. (2006). *Dictionnaire touareg de Mali*. Paris: Karthala

Huyghe, G. (1906). *Dictionnaire français-chaouia.* Alger: Lithographie Adolfe Jourdan.

Ibañez, F. E. (1949). *Diccionario rifeño-español.* Madrid: Instituto de Estudios Africanos.

Ibañez, F. E. (1959). *Diccionario espanol-senhayi*. Madrid: Consejo superior de investigaciones cientificas.

Kossmann, M. G. (1997). *Grammaire du berbère de Figuig (Maroc Oriental).* Louvain/Paris: Peeters.

Kossmann, M.G. (2000). *Esquisse grammaticale du rifain oriental*. Paris/Louvain: Peeters.

Kossmann, M.G. (2009). Tarifiyt Berber Vocabulary. In Haspelmath, M and Tadmor, U. editors, *World Loanword Database (WOLD)*, pp. 1533 entries. München: Max Planck Digital Library.

Lanfry, J. (1968), *Ghadamès: étude linguistique et ethnographique*. Alger: Fichier de documentation berbère.

Lanfry, J. (1973). *Ghadamès II. Glossaire (parler des Ayt Waziten)*. Alger: Fichier périodique.

Laoust, E. (1912). *Etude sur le dialecte berbère du Chenoua comparé avec ceux des Beni-Menacer et des Beni-Saleh*. Paris: Ernest Leroux.

Laoust, E. (1920). *Mots et choses berbères : notes de linguistique et d'ethnographie. Dialectes du Maroc*. Paris: A. Challamel.

Laoust, E. (1932). *Siwa: son parler*. Paris: Ernest Leroux.

Mammeri, M. (2003). *L'Ahellil du Gourara*. Paris: Éditions de la Maison des sciences de l'homme.

Naït-Zerrad, K. (1998). *Dictionnaire des racines berbères (formes attestées), I. A-BƐẒL*. Paris-Louvain: Peeters.

Naït-Zerrad, K. (1999). *Dictionnaire des racines berbères (formes attestées), II. C-DEN*. Paris-Louvain: Peeters.

Naït-Zerrad, K. (2002). *Dictionnaire des racines berbères (formes attestées), III. Ḍ- GƐY*. Paris-Louvain: Peeters.

Nehlil, -. (1909). *Etude sur le dialecte de Ghat*. Paris: Ernest Leroux.

Provotelle, P. (1911). *Etude sur la tamazir't ou zénatia de Qalaât es-Sened (Tunisie)*. Paris: Ernest Leroux.

Rahhou, R. (2005). *Dictionnaire berbère-français (parler des Beni-Iznassen)*. Thèse de Doctorat. Fès: Université Mohammed Ben Abdallah.

Renisio, A. (1932). *Etude sur les dialectes berbères des Beni Iznassen, du Rif et des Senhaja de Srair*. Paris: E. Leroux.

Saa, F. (1995). *Aspects de la morphologie et de la phonologie du berbère parlé dans le ksar Zenaga à Figuig*. Thèse de Doctorat. Paris: Université Paris III.

Sahli, A. (2008). *Mu□ jam □ amāzīghi □ arabi (ḫāṣṣ bilahajat □ ahālī fijīj)*. Oujda: Maṭābi□ al-□ anwār al-maghāribiya.

Serhoual, M. (2002). *Dictionnaire tarifit-français et essai de lexicologie amazighe*. Thèse de Doctorat d'Etat. Tetuan: Université Abdelmalek Essaâdi.

Stroomer, H. (1994). Morphological segmentation in Tachelhiyt Berber (Morocco). *Études et documents berbères*, 11: 91–96

Taïfi, M. (1992). *Dictionnaire tamazight-français*. Paris: l'Harmattan.

Taine-Cheikh, C. (2010). *Dictionnaire français-zénaga (berbère de Mauritanie), avec renvoi au classement par racines du Dictionnaire zénaga-français*. Köln: Rüdiger Köppe Verlag.

Yeou, M., Kiyoshi H. and Maeda S. (2011). Articulatory and acoustic characteristics of some geminates in Figuig Berber. *Proceedings of the 9th International Seminar on Speech Production*, Montreal.

Appendix: **example of dictionary page and finder list**

DN

adan (w-) *n.* ♦ intestins (en général), le petit intestin. ◊ **adanen n tmurt** litt. intestins de terre: lombric, ver de terre. cf. 'adan n tmuat' (Rif): m. sn. *Cf.:* **tameswadant, √MSWDN; imeḍren, √ḌṚN; tasuft, √SF**. *Pl.:* **adanen**. *dial & étym.:* adan, aḍan (Chl), adan (Mc, Izn, Rif, Sen, Cha, Che, Ace, Sns, Mzb, Oua, Ghd, Ght, Nef), adan, aden (Snd), ădân (Aha): boyau, intestins comme sn. commun; dér. du v. eden (Aha): graisser, être graissé.

tadunt (ḍḍ-/td-) *n.* ♦ graisse (d'origine animale). ◊ **mi ḍḍ un (ul) ṭṣṣeḥ ḍḍunt ukk ṃuṣ̌ṣ̌ iqqar (iqqaṛ) as yexx / mikk un (ul) teṛxis ḍḍunt ukk ṃuṣ̌ṣ̌ iqqar (iqqaṛ) ammu tfuḥ** litt. quand le chat n'arrive pas à avoir la graisse il lui dit "pouah!" / qu'est ce qu'elle pue!: se dit pour qqn. qui minimise l'importance d'une chose souhaitable, mais qu'il est incapable de réaliser (dicton). *syn.:* **tilebḍin, LBḌ**. *dial & étym.:* tadunt (Chl, Ntf, Sen, Mc, Izn, Rif, Cha, Sns, Mzb, Oua, Nef, Tim, Snd), tadunt, tadwent (Ace), tâdent (Aha), tadent, tedent (Nig): m. sn.

adun *n.* ♦ augment. du précéd.

DN

aden yuden, un (ul) yudin, ittaden, un (ul) ittiden *v.* ♦ couvrir, recouvrir, couvrir de couverture (dormeur); être couvert ◊ **adn it ammen ad ịṭṭeṣ** couvre-le d'une couverture pour qu'il dorme. ◊ **adfel qa□ illa yuden aḍraṛ** la neige a complètement couvert la montagne. ◊ **tella tuden imma nnes an (al) ixef nnes** elle s'est couverte jusqu'à la tête. ♦ fig. couvrir, protéger, chercher à innocenter (coupable, accusé). ◊ **itekk ẓẓbayel ppas ittaden xfes** il fait des fautes graves, mais son père ne le dénonce pas pour le protéger. *dial & étym.:* aden (Mzb, Oua, Izn, Ace, Sns, Tim, Snd, Ghd, Nef): couvrir et/ou pass.

ttwaden ittwaden, un (ul) ittwaden, ittwadan, un (ul) ittwidin *v.* ♦ être couvert, recouvert ◊ **tettwaden tmurt s wedfel (udfel)** la terre a été couverte de neige. ♦ fig. être protégé, innocenté, couvert

idan (y-) *n. vb.* ♦ fait de couvrir, de recouvrir, de couvrir de couverture; ♦ fait de chercher à innocenter.

madun (u-) *n.* ♦ plaque, dalle de pierre; ♦ dalle de tombeau. ◊ **i□emdan wala imudan** litt. il vaut mieux [s'appuyer sur] une canne qu'une dalle de tombeau: plutôt souffrir que mourir (dicton) *syn.:* **taḍelḥa, √ḌLḤ**. *Pl.:* **imadunen**. *dial & étym.:* madun, tmadunt (Mzb), teddenen (Zen): mm. sn.

tmadunt *n.* ♦ couvercle, tout ce qui couvre; ♦ bouchon. ◊ **tmadunt n uqlil** couverle de cruche, d'une théière. ◊ **tmadunt n uɣ ellay** couverle de bouilloire. ◊ **tmadunt n tmeṛmiṭṭ** couvercle de marmite. ◊ **tmadunt n qeṛ□et** bouchon de bouteille. ◊ **tmadunt n tiṭṭ** litt. couvercle de l'œil: paupière (supérieure). ◊ **yuf uqlil tmadunt nnes** litt. la cruche a trouvé son couvercle: qui se ressemble s'assemble (dicton). *Pl.:* **timadunin**. *dial & étym.:* madun (Tim), addan (Mzb), adan, badun (Oua), amaden (Ghd): m. sn.; madun (Izn, Rif), mudun (Ace): couscoussier.

---

dent *n.* **tiɣ mest (te-), ƔMS.**
dent (de clé) *n.* **tiswet (te-), SW.**
dent (de fourche) *n.* **qaššaw (u-), QŠ(W).**
depuis *prép.* **si, S.**
dernier *n. adj.* **anekkaṛu (u-), KṚ.**
dernier-né *n. adj.* **amaẓuẓ (u-), MẒ.**
derrière *prép. adv. n.* **deffer, DFR.**
descendre *v.* **hwa, HW;**
v. **zder, ZDR.**
descente *n.* **asehnennay (u-), HNY;**
n. vb. **hekku (u-), HW.**
dessous *n.* **aḍḍay (w-), DY.**
dessèchement *n. vb.* **asiqquṛ (u-), ƔṚ.**
dessécher *v.* **ssiqquṛ, ƔṚ.**
dessécher (se) *v.* **aḍen, ḌN;**
v. **hšišeṛ, HŠṚ.**
desséché *n. adj.* **maḍun (u-), ḌN;**
adj. n. **ameqquṛ (u-), ƔṚ.**
destinée *n.* **mimun (u-), MN.**
dette *n.* **ameṛwaṣ (u-), ṚWṢ.**
deuil *n.* **anebdi (u-), NBD.**
deux (fém.) *n. adj.* **sent, SN.**
deux (masc.) *n. adj.* **senn, SN.**
devancer *n.* **asul, SL;**
v. **zzar, ZR.**
devant *prép. adv. n.* **zzat, ZT.**
devenir *v.* **dwel, DWL;**
v. **ḍḥa**

Les ressources langagières pour la recherche d'information textuelle: Cas de la langue amazighe

Fadoua Ataa Allah[1], Siham Boulaknadel[1]

[1]CEISIC, IRCAM
{ataaallah, boulaknadel}@ircam.ma

Résumé

Le passage de la langue amazighe de l'orale à l'écrit lui a permis d'être doté d'un système d'écriture électronique assurant son intégration auprès de ses consœurs dans le domaine des nouvelles de l'information et de la communication. Néanmoins, cette intégration suscite l'élaboration aussi d'outils et de ressources langagières particulièrement pour la recherche d'information.

Dans ce contexte s'inscrit cet article qui consiste à décrire les différentes ressources langagières pour la recherche d'information et leur élaboration dans la perspective de les exploiter dans un système de recherche d'information dédié à la langue amazighe pour améliorer l'accès à l'information.

1. Introduction

La recherche d'information est une discipline qui a évoluée judicieusement dans le temps. Elle remonte au début des années 1950, où elle a été destinée principalement à étudier et concevoir des outils de recherche réservés à une communauté de spécialistes (Mooers, 1948). Particulièrement, les premiers systèmes de recherche d'information ont été construits afin d'aider les bibliothécaires à retrouver des documents contenus dans des bases bibliographiques. Or avec l'avènement du Web, la recherche d'information s'est vulgarisée notamment par le biais des moteurs de recherche. Ainsi, l'explosion des données numériques disponibles a rendu le recours à des moyens de recherche performants et automatiques indispensable.

De ce fait, la recherche d'information textuelle a évolué de la recherche documentaire proprement dite vers des tâches de plus en plus nombreuses et diversifiées permettant le stockage, l'analyse et la recherche de tout type de médias (texte, audio, image et vidéo). Néanmoins, les systèmes de recherche d'information suscitent toujours l'intérêt des chercheurs afin de minimiser l'ambiguïté que peut entraîner une langue. Cette ambiguïté peut découler des différentes formulations que peut avoir un même concept, où les documents pertinents contiennent des termes sémantiquement proches de ceux de la requête mais exprimés différemment par le biais de la synonymie, l'hyperonymie, la flexion ou de la dérivation. Comme, elle peut être à l'origine d'un problème de polysémie (Moreau et Sébillot, 2005).

Afin de relever ces défis liés à la complexité du langage naturel, plusieurs travaux basés généralement sur l'intégration de ressources langagières ont été réalisés (Agirre et al., 2010 ; Dolamic et Savoy, 2010 ; Guelfi et al., 2007 ; Moreau F., 2006). Ces travaux ont prouvé, à travers les évaluations effectuées et les statistiques réalisées à base des corpus et des méthodes d'évaluation bien choisis (De Loupy, 2001), que l'utilisation de ressources langagières a favorisé une meilleur représentation d'une part du contenu informationnel et d'autre part le besoin des utilisateurs, permettant ainsi l'amélioration des systèmes de recherche d'information.

Dans ce contexte, nous avons adopté au sein du centre des études informatiques, système d'information et de communication une stratégie progressive d'élaboration de ressources langagières dédiées à la langue amazighe, en vue de les exploiter pour améliorer son système de recherche d'information intégré dans le site officielle de l'IRCAM (Ataa Allah et Boulaknadel, 2010). Ainsi, nous avons dans un premier temps procédé par la construction selon des normes reconnues dans ce domaine des ressources électroniques basiques et élémentaires, à savoir la liste des mots anti-dictionnaires et le corpus d'évaluation, dans la perspective d'élaborer des ressources langagières plus avancées tel que les vocabulaires contrôlés et les thésaurus.

Dans la suite de cet article, nous définissons et nous exposons, dans la section 2, la structure d'un système de recherche d'information. Ensuite dans la section 3, nous proposons une classification pour les ressources langagières selon la structure d'un système de recherche d'information. Dans la section 4, nous présentons un descriptif des ressources exploitées dans la recherche d'information. Puis avant de conclure, nous exposons les ressources en cours de réalisation dédiées pour la langue amazighe.

2. Système de recherche d'information

Le système de recherche d'information consiste à établir une correspondance pertinente entre l'information recherchée, représentée généralement par le biais d'une requête, et l'ensemble des documents disponibles. Ainsi, le système s'appuie sur un ensemble de processus, articulés autour de deux étapes essentielles : la phase d'indexation des documents disponibles et la phase de recherche ou de l'interrogation du fonds documentaire ainsi constitué (Bonnel et Moreau, 2005 ; Benséfia et al., 2003).

La phase d'indexation consiste à analyser les documents et les requêtes afin d'extraire le jeu de descripteurs qui permettra la représentation de leur contenu textuel et leur exploitation par un modèle de recherche d'information prédéterminé. A la base de ce modèle, l'étape de recherche vise à apparier les documents et la requête de l'utilisateur en comparant leur représentation. Puis, de sélectionner et d'afficher les documents les plus pertinents, dont les descripteurs d'indexation sont les plus proches de ceux de la requête.

En outre de ces deux phases, le processus de recherche d'information peut intégrer une étape de reformulation afin d'améliorer la performance du système en tentant de rapprocher la requête de l'utilisateur de son besoin d'information initial. Or, une telle reformulation n'est adoptée qu'après une phase d'étude et d'évaluation.

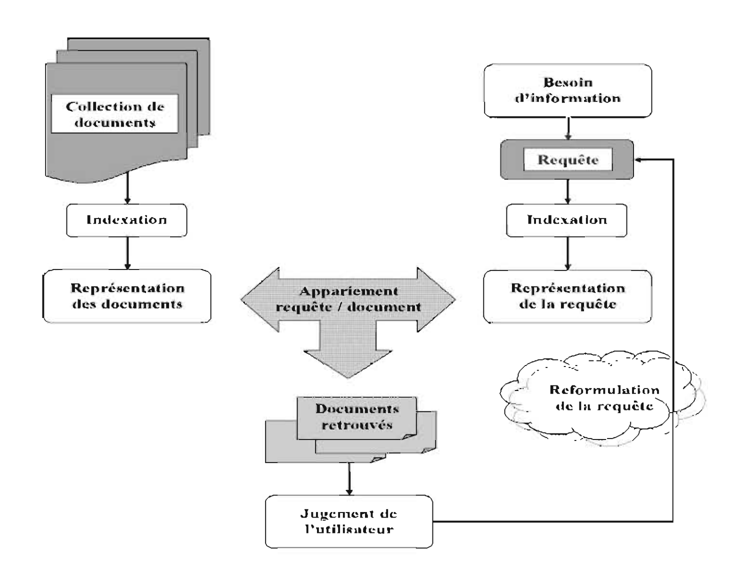


Figure 1. Structure d'un système de recherche d'information

3. Ressources langagières pour la recherche d'information

Dans un processus de recherche d'information, une grande variété de ressources langagières peut être exploitée. Ces ressources peuvent être classées selon leur intégration dans les phases de ce processus.

3.1. Phase d'indexation

Dans l'intérêt d'améliorer la performance des systèmes de recherche d'information, plusieurs types d'informations linguistiques peuvent être exploités pour enrichir les termes du jeu descripteurs d'indexation. Ces informations sont soient extraites à l'aide des outils linguistiques tels que les lemmatiseurs qui permettent de regrouper les mots selon une forme lexicale choisie par convention au sein d'un paradigme flexionnel, les racineurs qui consistent à regrouper les mots de la même famille selon leur racine, et les analyseurs morpho-syntaxiques qui se basent sur l'étude des formes et des règles de combinaison des morphèmes (Moreau et al., 2007); ou à partir de ressources langagières telles que les listes des mots anti-dictionnaires, les vocabulaires contrôlés et les ontologies (Brisaboa et al., 2010).

3.2. Phase de reformulation

Généralement, les utilisateurs des systèmes de recherche d'information, particulièrement les moteurs de recherche, ne sont pas des professionnels de la documentation. Donc il est difficile pour la plupart de ces utilisateurs de formuler idéalement leurs requêtes exprimant le mieux leurs besoins en terme d'informations (Smail, 1998). Afin de surmonter ce problème, les systèmes de recherche d'information souvent intègre une étape de reformulation automatique de la requête dans leur mécanisme. Elle consiste à modifier la requête initiale, en ajoutant de nouveaux termes extraits des résultats de recherches précédentes ou d'une base de connaissance telle que les thésaurus, les vocabulaires contrôlés et les ontologies (Hoang Diem, 2009 ; Mandal et al., 1998).

3.3. Phase d'évaluation

Avec l'évolution des techniques du traitement automatique des langues, l'évaluation de la performance des systèmes de recherche d'information est devenue une étape indispensable dans le processus de conception de ces systèmes. Elle sert à mesurer la pertinence des résultats vis à vis du besoin des utilisateurs. Dans ce contexte, les campagnes d'évaluation telles que TREC (Text REtreival Conference) et CLEF (Cross-Language Evaluation Forum) proposent des plates-

formes qui réunissent des protocoles d'évaluation et des collections de test volumineuses contenant des documents, des requêtes préalablement construites et des jugements de pertinence associés.

4. Descriptif des ressources langagières pour la recherche d'information

Parmi les ressources langagières les plus utilisées dans le domaine de la recherche d'information, nous citons les listes des mots anti-dictionnaires, les vocabulaires contrôlés, les thésaurus, les ontologies, et les corpus.

4.1. Liste des mots anti-dictionnaires

La liste des mots anti-dictionnaires consiste en un ensemble de mots ou de termes déterminés comme étant des mots peu informatifs et non pertinents pour la recherche d'information. Ces mots s'appellent aussi des mots vides, « grammaticaux » ou des mots outils. Généralement, ils se composent de prépositions, d'articles, de pronoms, d'auxiliaires ou encore de mots très fréquents au sein d'une collection de textes spécifique à un domaine donné.

4.2. Vocabulaires contrôlés

Un vocabulaire contrôlé, en sciences de l'information, est une liste de mots et d'expressions soigneusement choisis afin d'étiqueter les documents de manière à rendre leur repérage lors d'une recherche plus facile. Les vocabulaires contrôlés permettent de résoudre les problèmes liés à l'homographie, la polysémie et la synonymie, par une relation bijective entre les concepts et les termes acceptés, et réduisent l'ambiguïté inhérente au langage humain naturel, où différents noms peuvent être attribués à un même concept.

4.3. Thésaurus

Dans un thésaurus, le vocabulaire contrôlé est organisé sous forme d'un ensemble hiérarchique de termes clés représentant des concepts d'un domaine particulier. Cette hiérarchisation peut correspondre à une spécialisation, où un terme du vocabulaire est lié à ses descendants par des relations précises ; ou à un élargissement, où le thésaurus donne de l'information sur des sujets connexes et relatifs au terme.

4.4. Ontologies

En informatique et en science de l'information, une ontologie correspond à un vocabulaire contrôlé et organisé et à la formalisation explicite des relations créées entre les différents termes de ce vocabulaire, permettant de donner un sens aux informations. Ces relations sont, le plus généralement, organisées par un graphe et peuvent être de type sémantique ou de composition et d'héritage.

4.5. Corpus

Les corpus sont des collections de données sélectionnées et organisées selon des critères explicites pour servir comme un échantillon d'une langue donnée pour un traitement particulier, ou comme une référence pour fournir une information en profondeur. D'où la nécessité que ces corpus doivent être suffisamment représentatifs d'une manière qu'ils contiennent toutes les variétés pertinentes d'une langue et de son vocabulaire.

En général, les corpus sont caractérisés par la nature de la langue traitée et par le contenu. Ainsi, un corpus peut traiter une langue ou plusieurs langue comme peut traiter du texte ou du multimédia. Son contenu peut être sous forme de données brutes ou de données enrichies par des annotations grammaticales et morphologiques ou des informations sémantiques.

5. Ressources langagières en langue amazighe

Après la conception et l'élaboration d'un moteur de recherche basique pour la langue amazighe (Ataa Allah et Boulaknadel, 2010), l'amélioration de sa performance par l'exploitation de connaissances linguistiques est une démarche progressive. Elle s'est initiée par l'intégration d'une liste de mots d'anti-dictionnaires conçue à partir de « La nouvelle grammaire de l'amazighe » (Boukhris et al., 2008) et la réalisation en cours d'un corpus d'évaluation selon les protocoles adoptés par TREC.

5.1. Liste des mots anti-dictionnaires

La liste des mots outils que nous avons réalisée est composée de particules d'aspect, pronoms personnels autonomes, pronoms affixes du verbe direct, pronoms affixes du verbe indirect, particules d'orientation, particules de négation, pronoms interrogatifs, subordonnants, pronoms affixes du nom ordinaire, démonstratifs de proximité, démonstratifs d'éloignement, démonstratifs d'absence, pronoms prépositionnels, pronoms démonstratifs, pronoms interrogatifs, pronoms

indéfinis, prépositions, adverbes de lieu, adverbes de temps, adverbes de quantité, adverbes interrogatifs de quantité, adverbes interrogatifs de manière, conjonctions, morphèmes du pluriel, particules préverbales, particules prédicatives.

Type	Particule
Conjonction	ⴰⵎ « am » *comme*
Adverbes de lieu	ⴷⴰ « da » *ici*
Préposition	ⵖⵓⵔ « γur » *chez*
Subordonnant	ⴰⵇⵇⴰ « aqqa » *que*
Pronoms indéfinis	ⵉⵛⵜ « ict » *une*

Table 1. Exemple de mots anti-dictionnaires

5.2. Corpus

Dans la perspective d'élaborer un corpus électronique d'évaluation pour les systèmes de recherche d'information de la langue amazighe, nous visons à construire un corpus représentatif qui contient des textes de différents genres, à savoir les textes scientifiques, médicaux, des articles journalistiques, des textes des sciences humaines, de contes et de poèmes. Or vu la rareté des textes écris en amazighe, nous sommes actuellement en cours de collecter au moins l'existant dans l'attente d'enrichir ce corpus dans le futur.

Par ailleurs, nous avons opté de suivre le protocole entrepris par les campagnes d'évaluation TREC (Craswell et al., 2005 ; Voorhees, 2005), afin que notre corpus soit élaboré selon une norme reconnue dans le domaine de la recherche d'information. Ainsi après la phase de la collecte, nous entamons une deuxième phase de dé-balisage des sources HTML qui sera succédée par une troisième étape qui consistera à convertir tous les textes en tifinaghe Unicode. Ensuite, nous regroupons tous les textes dans un seul fichier, où chaque texte est étiqueté par son identificateur, sa date de publication, sa catégorie et son contenu comme s'est représenté dans la table 2.

```
<doc>
<docid> 10 </docid>
<date>2008 </date>
<fld> story </fld>
<text>
<title>
ⵉⵏⵙⵉ ⴷ ⵡⵓⵛⵛⵏ
</title>
ⵢⴰⵏ ⵡⴰⵙⵙ, ⵉⴼⴼⵖ ⵃⵎⵎⵉ ⵉⵏⵙⵉ, ⵙⴳ ⵜⴰⴷⴷⴰⵔⵜ ⵏⵏⵙ, ⵉⴷⴷⴰ ⵖⵔ ⵜⴰⴳⴰⵏⵜ
ⴰⴷ ⵉⵙⵙⴰⵔⴰ.
ⵢⵓⴼⴰ ⴳ ⵜⴰⴳⴰⵏⵜ ⵃⵎⵎⵉ ⵓⵛⵛⵏ.
ⵉⵏⵏⴰ ⴰⵙ ⵢⵉⵏⵙⵉ: … .
</text>
</doc>
```

Table 2. Exemple de document

Après, nous construisons un jeu de requêtes en s'inspirant toujours des campagnes d'évaluation TREC. Ces requêtes comportent quatre champs : un identificateur, un titre nommant le thème ; une description énonçant l'objet de la recherche ; et un développement explicitant des critères de validité des rapprochements, des mots-clés fournissant le contexte terminologique et les concepts concernés. Cette forme apporte une information aussi complète et détaillée que possible, y compris des connaissances avancées sur le domaine grâce aux mots-clés. Un exemple de ces requêtes est présenté dans la table 3.

```
<top>
<num> Number: 2 </num>
<title>
ⵜⴰⵔⵡⴰ ⵏ ⵜⵎⵓⵔⵜ
</title>
<desc>  Description :
ⵔⵣⵓ ⴳⴼ ⵓⴹⵕⵉⵚ ⵉⵙⵙⴰⵡⴰⵍⵏ ⴳⴼ ⵜⴰⵔⵡⴰ ⵏ ⵜⵎⵓⵔⵜ
</desc>
<narr> Narrative:
ⵉⴹⵕⵉⵚⵏ ⵉⵙⵙⴰⵡⵍⵏ ⴳⴼ ⵓⵎⵢⴰⵙⵙⴰ ⵊⴰⵔ ⵜⴰⵔⵡⴰ ⵏ ⵜⵎⵓⵔⵜ ⴷ ⵜⴰⵔⵡⴰ ⵏ
ⵓⴳⵏⵏⴰ, ⵜⴰⵢⵔⵉ ⵏ ⵜⵎⵓⵔⵜ, ⵜⵓⴷⵔⵜ ⵜⴻⵜⵜⴼⵖ ⴷ ⵙⴳ ⵓⵃⴷⴷⵉⵙ ⵏ ⵜⵎⵓⵔⵜ.
</narr>
</top>
```

Table 3. Exemple de requêtes

Puis, nous constituons une liste de jugement de pertinence qui se base sur une sélection manuelle, par plusieurs locuteurs de la langue amazighe, des documents pertinents pour chaque requête.

En outre, au cours de l'élaboration de notre corpus, nous avons envisagé qu'il vérifie trois types de conditions : les conditions de signifiance, les conditions d'acceptabilité, et les conditions d'exploitabilité (Pincemin, 1999).

- Conditions de signifiance : notre corpus est constitué en vue de l'évaluation de la performance des améliorations que nous désirons apporté au système de recherche d'information que nous avons élaboré pour la langue amazighe. Les documents retenus sont extraits des différents ouvrages édités par notre institut ainsi que des pages Web du son site officiel.

- Conditions d'acceptabilité : le corpus apporte une représentation fidèle, sans aucune modification au niveau du contenu à l'exception de changement du codage utilisé en cas de besoin, de manière à avoir une homogénéité au niveau du codage et de ne garder que l'UNICODE. Par ailleurs, le modèle de requête utilisé assurera un niveau de détail adapté au degré de finesse et à la richesse attendue en résultat de l'analyse.

- Conditions d'exploitabilité : nous envisageons que les textes qui formeront notre corpus seront commensurables de telle sorte que leur contenu représente l'intégralité de l'ouvrage ou de la page Web exploités. En outre, nous visons que ce corpus devrait apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs (au sens statistique du terme).

6. Conclusion

Le manque des ressources langagières pour la langue amazighe a un impact majeur sur l'amélioration d'accès à l'information et sur la diffusion de la langue amazighe à travers le Net. Ainsi dans la perspective de surmonter cet obstacle, nous avons envisagé dans cet article d'énumérer les différentes ressources utiles pour la recherche d'information et de décrire la méthodologie entreprise dans la réalisation des ressources langagières basiques.

Références :

Agirre E., Arregi X., Otegi A. (2010). Document Expansion Based on WordNet for Robust IR. Actes de Coling'2010 (23rd International Conference), pp. 9-17. Beijing, China.

Ataa Allah F. et Boulaknadel S. (2010). Amazigh Search Engine: Tifinaghe Character Based Approach, *Actes de IKE'2010*, pp. 255-259.

Benséfia A., Paquet T. et Heutte L. (2003). Documents Manuscrits et Recherche d'Information, *Revue Document Numérique*, vol. (7) : 47-60.

Bonnel N. et Moreau F. (2005). Quel avenir pour les moteurs de recherche ?. *Actes de MajecSTIC '05*.

Boukhris F., Boumalk A., Elmoujahid E. et Souifi H. (2008). *La nouvelle grammaire de l'amazighe*, IRCAM, Rabat, Maroc.

Brisaboa N.R., Luaces M. R., Places A. S. et Seco D. (2010). Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index. *GeoInformatica*, vol. (14) : 307-331.

Craswell N., Hawking D., Wilkinson R., et Wu M.(2005). Overview of the TREC 2004 web track. *Actes de TREC 2004 (13ème Conférence)*.

De Loupy C. (2001). L'apport de connaissances linguistiques en recherche documentaire. *Actes de TALN'2001*, Tome 2, pp. 129-133. Tours, France.

Dolamic L. et Savoy J. (2010). When stopword lists make the difference. Journal of the American Society for Information Science and Technology, vol. (61): 200–203.

Guelfi N., Pruski C. et Reynaud C. (2007). Les ontologies pour la recherche ciblée d'information sur le Web : une utilisation et extension d'owl pour l'expansion de requêtes. *Actes d'IC'2007 (18ème journées francophones)*, Grenoble, France.

Hoang Diem L. T. (2009). Utilisation de ressources externes dans un modèle Bayésien de Recherche d'Information. Application à la recherche d'information multilingue avec UMLS. Thèse de Doctorat, Université de Joseph Fourier, Grenoble, France.

Mandal R., Takenobu T. et Hozumi T. (1998). The Use of WordNet in Information Retrieval. *Actes de COLING/ACL'98*, pp. 469-477.

Mooers C.N. (1948). Application of Random Codes to the Gathering of Statistical Information. Thèse de Master, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Moreau F. (2006). Revisiter le couplage traitement automatique des langues et recherche d'information. Thèse de Doctorat, Université de Rennes 1, Rennes, France.

Moreau F., Claveau V. et Sébillot P. (2007). Intégrer plus de connaissances linguistiques en recherche d'information peut-il augmenter les performances des systèmes ?. *Actes de CORIA'07 (4ème Conférence).*

Moreau, F. et Sébillot, P. (2005). Contributions des techniques du traitement automatique des langues à la recherche d'information. Rapport de recherche n° 1690, IRISA, Rennes, France.

Pincemin B. (1999). Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative. *Actes de TALN'99 (Atelier Corpus et TAL : pour une réflexion méthodologique)*, pp. 26-36.

Smail M. (1998). Vers des systèmes évolutifs de recherche d'information : un état de l'art. *Technique et Science Informatiques*, vol. (17): 1193-1222.

Voorhees E.M. (2005). "Overview of TREC 2004". *Actes de TREC 2004 (13ème Conférence).*

A universal Amazigh keyboard for Latin script and Tifinagh

Paul Anderson

p_j_anderson@volny.cz

1. Introduction

Systems of Amazigh text encoding and corresponding keyboard layouts have tended to be narrowly aimed at specific user communities, because of differences in phonology and orthography across Amazigh language variants[1].

Keyboard layouts for language variants have therefore lacked orthographic features found in other regions. This restricted focus impedes users' experimentation with the writing of other Amazigh regional variants and converged literary forms where they differ in orthographic features or in script. So far there has been no way to type more than a handful of Amazigh variants intuitively on any one layout even within one script.

This fragmented development has meant that keyboard driver implementations have often lagged behind advances in technology, and have usually failed to take into account general keyboard layout design, ergonomy and typing speed, and solutions from other Amazigh regions or non-Amazigh languages. Some users even preferred to improvise key definitions based on their own understanding, which often resulted in mistaken use of lookalike letters and diacritics.

Keyboard layouts have also failed to provide for Amazigh minority populations around the world, and have considered the multilingual context of Amazigh language use only locally.

Several scripts are commonly used to write Amazigh variants, and even within a script there are different orthographies in use. Some orthographies are formal

---

[1] I use the term 'language variant' since distinguishing 'dialect' and 'language' is not necessary here.

standards. In others, some features are obsolete but still in use, some features are still disputed, and some features are regional usages or personal initiatives, or are required only for writing more phonetically. Complete descriptions of phonology and orthographies are often difficult to find. It is therefore complex to determine, for each script, a sufficient and practical superset of features for writing a large set of language variants so that keyboard layouts can be harmonised.

This project began because I was creating a Kabyle dictionary document for my own use and existing Amazigh keyboard layouts did not produce a suitable set of letters.

Keyboard layout design ties in closely with Unicode encoding, fonts, and font rendering capabilities of software. To facilitate good design, the project spawned a separate but related investigation into the possible ways of encoding Tifinagh text based on the existing Tifinagh set in Unicode, and their effectiveness in representing different regional Tifinagh repertoires and orthographies. The results of the investigation (Anderson, 2010a) were submitted to the Unicode consortium and considered by the Technical Committee in late 2010. The investigation also led to two of the letters that were presented but deferred in the original Tifinagh Unicode proposal (Andries, 2004) being proposed (Anderson, 2010b) and accepted into the encoding process by the consortium.

Quality Amazigh keyboard layouts would allow easy production of well-encoded text. Their widespread use, with fonts of equivalent standard, would promote good document production without mistaken use of look-alike characters or diacritics, and stimulate creative output. Quality layouts would also promote the use of Unicode, consistent with other languages, and show a solid base in technology for Amazigh, boosting its prestige.

Further, if all regional keyboards could be used to type converged literary forms of Amazigh, there would be no technical barrier to prevent experimentation and adoption by potential users. Also, if keyboards could be used to type many regional forms, writers of a variant would easily be able to type it correctly even in a place that used a different standard orthography. It would be easier to become familiar with other variants.

Freely available keyboard layouts (and fonts) that were reusable across Amazigh variants would allow resources to be pooled to achieve high quality more quickly. These tools would form a stable foundation for work in other areas of technology for the language and in language maintenance.

This project's results are intended to fulfil those needs.

Viewing the scripts and all the varied orthographic solutions within them as writing tools to be evaluated and adapted across all Amazigh variants, encourages technical evolution of the writing systems, and also creativity in Tifinagh typography. Local traditions become simply styles of writing Amazigh, rather than constraints. The Tifinagh script and its future belong entirely to writers of Amazigh.

2. Aims

The project targeted typing Northern Amazigh in Latin script across Morocco and Algeria, and typing as many major Amazigh variants as possible in vowelled Tifinagh (to provide support for the latest orthographic advances). Prioritisation of Amazigh variants was by level of representation in modern Amazigh literature and by whether they are currently written in either the Latin script or Tifinagh. Arabic script was judged out of scope[2], as an ordinary Arabic keyboard can be used, and Arabic Amazigh orthography is not official or prevalent (though it has significant representation in modern literature in Morocco, on Algerian state television and some official Algerian websites).

Here, I use the loose term 'Northern Amazigh' to group language varieties having one short vowel and three long vowels, distinguishing them from the 'Tuareg' varieties with their richer vowel repertoire, while recognising that varieties such as Siwi and Ghadamsi evade these categories.

As well as local Amazigh variants, the project considered Northern Amazigh as a whole, targeting the superset of orthographic features needed for both Ircam's standard Amazigh and possible future converged forms. Similarly, Tuareg was considered as a whole. The project also examined the extent to which orthographies and text encoding could be shared across all Amazigh variants.

The Latin style used in Algeria for Tuareg transcription was a priority, to cover Algerian needs also for Tuareg. Another priority was to include experimental features to allow a Latin transcription even more consistent with Northern Amazigh orthography. Support for West African-style Latin script for Tuareg was only a secondary aim because the Tuareg zone has Amazigh variants as recognised

---

[2] Versions of the project's Latin and Tifinagh layouts adapted to match Arabic keytops might however be a useful future development for those used to Arabic keyboards or Arabic Amazigh orthography.

national languages and there is official support for them in Latin script (but not for Tifinagh). Provision for typing Tuareg Tifinagh in classical style with limited vowel marking and with ligatures was similarly desirable but non-essential.

Other Amazigh variants were to be covered for Latin and Tifinagh to the extent that information was available, but not necessarily for typing intuitively (letters could be fitted in ad hoc), and for transcription rather than practical use where there was no local Latin or Tifinagh writing tradition.

The project aimed to provide drivers for keyboard arrangements covering the writing of the targeted Amazigh variants in Tifinagh and Latin. Each arrangement, for a set of variants, was to enable a complete set of orthographic features for a script – hence 'universal' – while remaining intuitive for typing its supported variants.

If possible the letters were to be laid out similarly for different regions and scripts, so that users could type different regional forms, in either Latin or Tifinagh, without confusion, but with the Tifinagh layout remaining natural for Tifinagh and the Latin layout natural for Latin. The ideal was a single arrangement per script sufficient for typing all targeted Amazigh variants, and intuitive enough to be preferred for that script by users.

One secondary aim was provision of obsolete features, to encourage users to adopt the new keyboard layouts and learn to bring their writing up-to-date. Another aim was to provide the ability to mix non-Amazigh languages. Another was to ensure that keyboard driver installation provides both Latin and Tifinagh capability together, for widest usability and to make it easy for Latin script users to try typing Tifinagh.

The layouts were to target primarily the French AZERTY physical keyboard found across North Africa, but also to contain intuitive adaptations for other physical keyboards used in countries outside Africa with significant Amazigh populations. All adaptations were to have equivalent Amazigh capabilities, so that any supported orthography could be typed on any adaptation.

Compatibility with Ircam's Tifinagh keyboard was a priority as an established standard.

Windows and Linux were to be the initial target platforms, in that order. Windows installations are widespread and familiar to users, and were the primary target. Linux is easy to contribute to, free, known in North Africa, and likely to grow in use there as technical knowledge of it increases, so it was the secondary target.

Apple is significant in publishing, and in mobile devices. Consideration of implementations on Apple products was left to future work.

3. Method

My approach was to look at the phonology across the language variants, and then find a minimum set of orthographic features in the Latin and Tifinagh scripts that constituted a consistent practical writing system across language variants in each script (with helpful written communication from Maarten Kossmann, Leiden University ; Lameen Souag, SOAS, University of London). I supplemented these Latin and Tifinagh character sets with additional features needed for regional orthographies where they could not be encoded in the shared way, as well as features for explicit phonetic writing, and obsolete, rare and disputed features.

For Tuareg Tifinagh the main priority was given to vowelled Tuareg neo-Tifinagh[3] orthography as the most up-to-date orthography. Several possible vowelling techniques (Elghamis, 2004 ; Louali, 1993 ; Issouf, 2007), none of which has achieved prevalence, had to be prioritised for support. The technique of the Association for the Promotion of Tifinagh, Niger (APT) was consistent with Northern Amazigh's marking of vowels, so it could be used together with any Tifinagh repertoire. Also, its extra letters, derived from traditional vowel letters (Amessalamine Ahmed, APT, written communication ; Elghamis, 2010), were graphically simple. Other techniques, while sometimes possible to encode in existing Unicode, would need a different keyboard layout, or special fonts to emulate the combinations of letters and diacritics (not good practice, but a possible temporary arrangement). SIL International's Tifinagh vowelling system, for example, was not consistent with Northern Amazigh practice, and though it could arguably be written in current Unicode after the addition of one APT letter, it would need a different keyboard layout to be typed practically (C. Grandouiller, L. Priest, J. Coblentz, SIL International, written communication).

Traditional Tifinagh styles with limited or no vowelling were to be supported too if possible, with traditional ligatures, as well as Tifinagh-specific punctuation.

---

3 I use the term 'neo-Tifinagh' to mean recent extended adaptations of Tifinagh with new and modified letters ; 'Northern' repertoires introduced from scratch, with letters from several historical sources, and 'Tuareg' repertoires more closely based on existing regional practice.

Phonetic features such as marking of Tuareg vowel nasalisation and consonant palatalisation were another secondary priority.

I then mapped this set of features onto Unicode. For Tifinagh, this required the addition of two APT letters needed for Tuareg's richer vowel repertoire, representing the Tuareg long vowels e and o, to the Unicode character set (they have been accepted for inclusion). Unicode's Tifinagh Joiner character was to be typable on the keyboard for generation of ligatures, as well as the Tifinagh Separator punctuation character.

The original encoding of the Tifinagh script in Unicode (Andries, 2004) was a major step forward for the script. However, Unicode has still not been adopted for Tifinagh in some regions where regional letter sets were incompletely encoded, or where font technology was inadequate until recently for rendering text correctly.

Analysis of the Tifinagh script encoding in Unicode (Anderson, 2010a) resulted in two possible ways of thinking about and using Unicode for Tifinagh. One was glyph-based, where each variant glyph is allocated a code point, with the result that each regional Amazigh variant would need different optimised tools for the same purpose. The other was letter-based, where variations in the symbol used to write a sound are left to the font and a smaller set of code points is used. The letter-based principle was proposed early in the original Tifinagh discussions (Everson, 1998) but has been diluted since by addition of code points for glyph variants.

The analysis concluded that glyph-based encoding was insufficient to encode Amazigh orthographic variants side by side in the same font, and would require further additions in any case. The letter-based encoding, by contrast, would already be near-complete. It would allow Amazigh variants to be encoded for Tifinagh with the same small shared set of code points, meaning that they could be typed using the same keyboard layout – and that any Tifinagh font could render different Amazigh variants in a consistent style. According to the Unicode technical committee, there is no technical objection to a letter-based interpretation and how the script is used in Unicode is up to the community.

The code point set to be generated by this single Tifinagh keyboard layout was hosen so that the standard reference glyphs shown by generic fonts would all follow Ircam's quality criteria for legibility (Ameur et al., 2006 ; Bouhjar, 2004), and the subset for the Moroccan alphabet would appear as Ircam's letter forms. This implies a change of encoding for Tuareg. Where the Tuareg letter form and the equivalent Ircam one differ, the Ircam letter's code point is used – coding letters by equivalence rather than visual appearance.

For the Latin script, the Latin rather than Greek form of gamma was chosen, as it can not be confused with y, has matching lower and upper case forms, and is becoming prevalent on the Internet (Brugnatelli, 2002). Similarly, the Latin form of open e was chosen rather than the Greek epsilon and sigma. The Greek forms were given reduced priority as obsolete but were retained for their users and also as calligraphic alternatives.

The letter representing ḇ (spirant 'b') was to be made easy to type, if possible, for both Latin and Tifinagh. It might be a viable alternative to 'v' for distinguishing 'b' from its spirant in writing for some proponents of this distinction in Kabylie, Algeria. Writing 'ḇ' would preserve orthographic coherence with other language variants.

Extra letters and punctuation were to be provided for occasional non-Amazigh language words in multilingual environments. To find this set, the main countries with Amazigh minorities were determined (including West Africa), and the national languages determined. The writing systems were then examined for features required (partly by examining keyboards existing for them), and the countries' physical keyboards were targeted for layout adaptations. Transcription of Arabic, especially Darja Arabic, was desirable, for the same purpose.

The next step, for each script, was the design of keyboard layouts for typing the set of orthographic features identified. Ergonomic considerations such as touch typing were taken into account. Prioritisation was needed in order to fit all the necessary letters and symbols into the limited space. Lower case letters and common punctuation were made the easiest to type, followed by capital letters, rarer letters and diacritics, then common symbols for programming, and finally rarer symbols. Letters were arranged by sound and shape similarity, using techniques similar to keyboards for other languages with similar requirements (such as multiple diacritics). The aim was to allow typing with as few keypresses as possible, as intuitively as possible. The keyboard was to produce correct Unicode sequences.

The AZERTY layouts were to form a simple learning path from the French AZERTY keyboard. Also, Amazigh functions were to be duplicated to use the same keys as existing Amazigh keyboards in common use, where possible. Obsolete orthographic features were to be harder to access than modern equivalents, though in intuitive positions if commonly used. These measures were to avoid confusing users and to maintain their productivity.

Bilingual Amazigh-French versions of the keyboard layouts were designed as a handy utility for users who frequently mix languages, or who prefer to adapt gradually to typing Amazigh by using a layout very close to what they are used to.

The keyboard layouts specific for Amazigh had additional design considerations. They were designed to be implemented cross-platform, following the constraints of Linux keyboard mappings and X input methods. Also, the layouts were to be first prototypes for future standard extensions to operating systems, or even for future inclusion as standard, for easy availability. For this reason they were designed to be adaptable for standards and future trends, for example by splitting Latin and Tifinagh into separate drivers, or by using Unicode combining diacritics instead of using deadkeys, or by moving secondary letters and diacritics to a separate layer (as in the Canadian Multilingual Standard keyboard and ISO/IEC 9995 - either the existing European-style secondary layer, or a new, African one, depending on the keyboard's region).

All layouts were designed to function whether installed on the French locale, or on future Amazigh locales (to obtain suitable spellchecking, autocorrection and other tools).

Physical keyboards commonly available in North Africa often lack a <> key or have the <> key or the * key moved to various locations. The layout design took this into account.

After design, the layouts were created for Microsoft Windows using Microsoft Keyboard Layout Creator (MSKLC) 1.4.

Any keyboard features made necessary by technical limitations of Windows, MSKLC or Microsoft Word were designed not to disturb the overall arrangement of keys.

For the forseeable future, until Amazigh layouts are bundled as standard with operating systems, layouts will be downloaded by users or installed from media. To ensure that Latin and Tifinagh capabilities are always installed together, they were implemented in the same keyboard layout, as MSKLC generates the installer automatically.

Prototype OpenType Tifinagh fonts (based on Hapax Berbère, with thanks to Patrick Andries) were created with fontforge under Linux, then additional tables were added with the Microsoft VOLT tool. Existing Ircam fonts remained compatible with the project's keyboard layouts but lacked non-Moroccan features, for example for writing Tuareg.

For each regional profile of keyboard use, the prototype fonts were used to check that existing font technology could give the correct regional appearance to the shared encoding, including all orthographic features. Software support was tested with MS Notepad, MS Wordpad, MS Word 2007, OpenOffice, and MS Word 2010 for advanced font features, on MS Windows XP service packs 2 and 3 and MS Windows 7.

Research about phonology and orthographies was examined, along with examples of literature, web sites, signage, and printed materials, as well as feedback on usability from users (special thanks to Kamal Bouamara, Bejaia University). This was done in iterations, with releases on the internet at each stage.

4. Results

The project has produced a set of keyboard layouts, currently available for Windows, freely downloadable from http://www.akufi.org/ . For the French AZERTY keyboard, there are adaptations for different levels of user aptitude - the Amazigh-centric layout (figs. 1a, 2a), and the Amazigh/French bilingual layout with minimal changes from standard French AZERTY (figs. 1b, 2b). There are adaptations of the Amazigh-centric layout suitable for several countries' physical keyboards, to keep letters and major punctuation in intuitive positions.

Current implementations have both Latin and Tifinagh in one driver, with the mode switched by Caps Lock (so capital letters are accessed only via Shift). Two versions of each driver are available, one defaulting to Latin and the other to Tifinagh. Regions can use the version that defaults to their official or prevailing script.

Diacritics are added in Latin mode by deadkey and in Tifinagh mode by typing Unicode combining diacritics after a letter.

The non-breaking hyphen is useful for ensuring that hyphenated constructions such as Kabyle 'iman-is' are never split between lines of text, and it is easy to type.

Tifinagh support for Tuareg is based on appropriate font rendering of a small set of code points in Unicode based on the letters for the Ircam alphabet, not on the Tuareg variant letters.

For users with legacy systems with poor rendering of advanced font features, temporary extended keyboard versions were created for typing presentation forms explicitly using the Private Use Area of fonts.

There is a provisional set of scientific characters available via AltGr+X.

An academic transcription feature gives users a quick alternative option to switching to a dedicated layout. It is a provisional and temporary utility, present as an extended version of the facilities of the INALCO 'unicber' keyboard, and not intended for future standard layouts.

Being essentially configuration files, automatically compiled into self-contained code modules, the layouts were licensed with the Creative Commons licence. The terms chosen require only acknowledgement of the author, so that use of the work is encouraged.

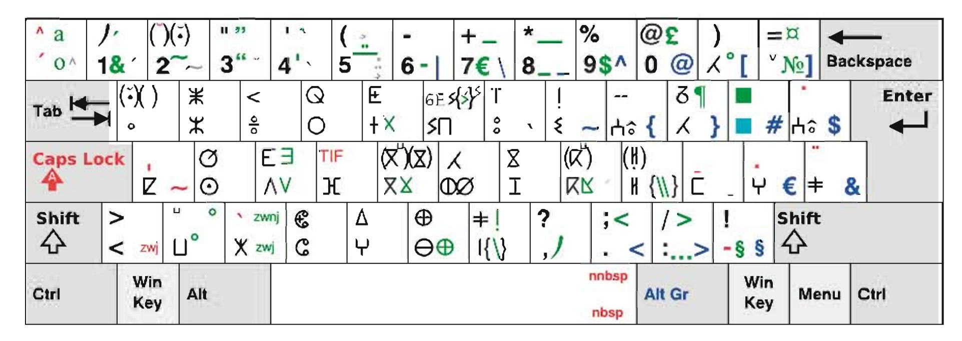


Figure 1a* : Amazigh-centric layout – Tifinagh

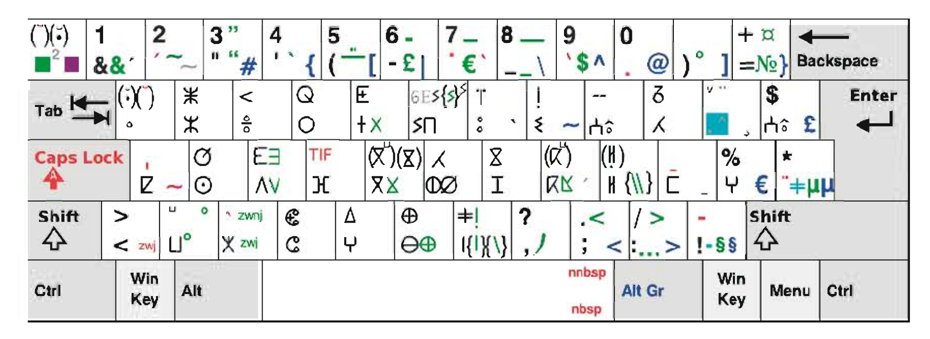


Figure 1b*: Bilingual Amazigh/French layout - Tifinagh

* Diagrams created with a generic template from Wikipedia (under Creative Commons BY-SA 3.0)

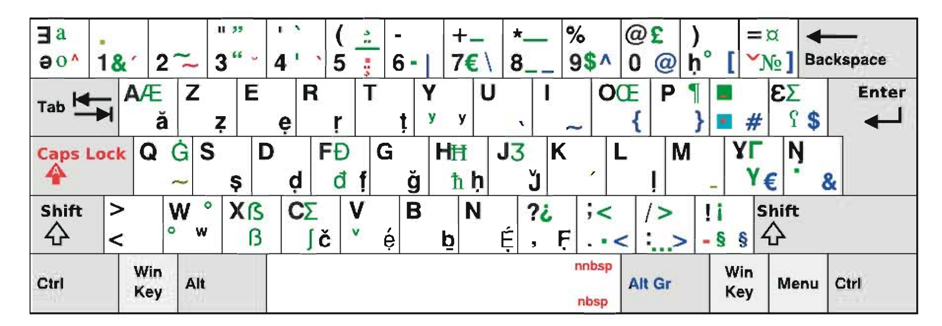


Figure 2a* : Amazigh-centric layout - Latin

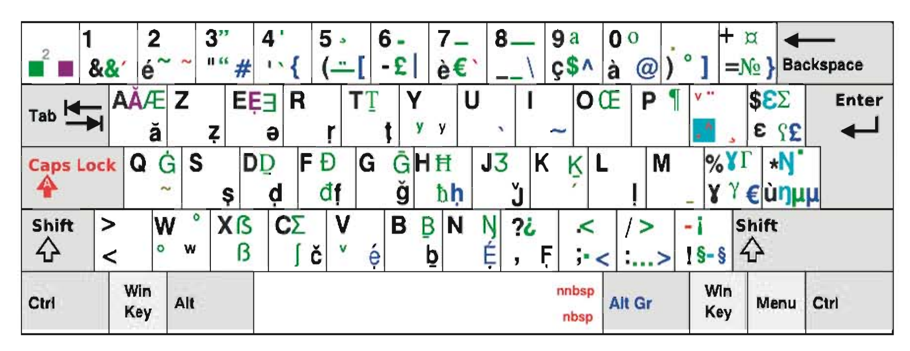


Figure 2b* : Bilingual Amazigh/French layout - Latin

5. Discussion

For Northern Amazigh : Moroccan Tarifit, Tamazight and Tachelhit, Algerian Tacenwit (Bouridj and Nouh, 2009), Kabyle, and Chaoui are supported.

For Tuareg : The forms of Algeria, Niger, and Mali are supported, though the only Tuareg neo-Tifinagh vowel system supported in a Unicode-compliant way is that of APT Niger. Writing of extra phonetic features in Tifinagh is partially supported (using substitute letters and techniques such as the Unicode private use area), with improvements left to future work.

The Tuareg Tifinagh orthography for ŋ and ɲ is still unstable (Elghamis, 2010). The candidate letters are supported, and can be typed in different Unicode forms. Refinement is left to future work. Also, dj-like sounds are documented in Tuareg

---

* Diagrams created with a generic template from Wikipedia (under Creative Commons BY-SA 3.0)

(e.g. Sudlow, 2001 ; Heath, 2005), and support for them may need to be added if existing mechanisms are insufficient.

The West African Latin orthographies of Tuareg are supported, including letters whose current use is uncertain.

Full analysis of Algerian oasis forms, Libyan and Egyptian forms was left to future work, but a first effort at support for Mozabite (Delheure, 1984), Gourara Amazigh (Bellil, 2006), Nefusi, and Siwi (Christfried Naumann, Institute for African Studies, Leipzig University, written communication), has been made.

Transcription of Zenaga, Tunisian Amazigh forms, and Ghadamsi remains to be assessed. Burkina Faso-style Latin orthography for Tuareg was left to further work in the absence of definitive Unicode information for emphatic consonants, but the layouts have suitable places available.

In future, on a machine installed in Tifinagh, in order to see their preferred Tifinagh letters, sers would have to set the system font. Currently, for use in applications only, users need to set the font used by their applications, for example for chat. For this reason the reference glyphs of the Ircam set, displayed as the default, are likely to become well known for use in informatics. Commonly used web sites such as for email do not generally allow font changes and not all browsers allow overrides. Confusion is unlikely because the Ircam equivalents of Tuareg letters do not resemble other Tuareg letters.

The layout for a script when it is non-default (i.e. accessed by pressing Caps Lock) has some small differences because of technical limitations in Windows, such as having to add a diacritic to a letter after typing it, or finding a combined letter and diacritic in a different key position. These differences (marked with parentheses on figs. 1a and 1b) have been made as unobtrusive as possible.

The Latin layout is not exactly symmetrical with the Tifinagh layout because the two scripts have slightly different arrangements which are natural to them. However, when a script is not the default and Caps Lock is used to access it, this secondary layout is slightly biased to be more intuitive for users of the default script. These differences are seen with ǧ, g^w and k^w, for example. A position on the 'j' key is intuitive for a Moroccan Tifinagh user because of the use of the Tifinagh equivalent of 'dj' or 'jj' for 'ǧ' in Moroccan Tifinagh orthography (Ameur et al., 2006 ; Ameur, 2004). In future the Tifinagh letter YADJ may be duplicated at AltGr+g and the Latin letter ǧ duplicated on AltGr+j for better symmetry.

Support for the Ircam Tifinagh keyboard's use of the 'o', 'p' and 'v' keys also disturbs the Latin/Tifinagh symmetry, but these letters are rare in Northern Amazigh and the equivalent Tifinagh letters are easily accessible via shift. For Tuareg, 'o' is more common, but both Tifinagh letter YE and Tifinagh letter YO are accessed via shift so it becomes natural to the user.

Since Tifinagh fonts can be reused between language variants, typographic creativity is opened up. Users can choose whether to use filled or closed dots, and whether sequences of Tifinagh letters YAL and YAN are disambiguated by linking lines in the letter forms, Moroccan-style, or by leaning or offsetting the letters (Algerian and Tuareg-style). Future development of cursive fonts, perhaps based on the work of Tuareg Tifinagh calligraphers, would have maximum utility. For orthographies other than classical Tuareg, fonts can provide ligatures purely for cosmetic effect, and the keyboard layouts allow control of ligature formation with the Zero Width Joiner character.

With operating system and software support, correct Unicode properties, and suitable fonts, Tifinagh could be written vertically like Japanese, or right-to-left, for typographic effect.

Historical Tifinagh or Libyco-Berber letters, excluded from consideration by the project, could be mapped to their equivalent letters in the code point set to the extent that the phonology allows, typed with a normal keyboard layout and rendered by special fonts.

That calligraphic Tuareg-style and historical fonts for Mozabite have been requested by one user of the project's keyboards shows the strength of the project's approach.

6. Conclusions

The project is unusual because its analysis was of both Latin and Tifinagh across each Amazigh variant. Also, the range of Amazigh variants included was international, not national or regional as had been the case for keyboards previously. In addition to phonological and orthographic descriptions, many examples of written language were studied for their typography.

The resulting keyboard layouts are 'universal' in that they can be used to type a wide range of Amazigh variants as well as converged Amazigh literary forms, all still with an intuitive, non-arbitrary arrangement of keys. Amazigh variants that had no previous support with Latin and Tifinagh can be transcribed. The

universality has a limitation – that only the APT[4] technique of Tuareg vowel marking is supported. This does not prevent encoding of other vowel mechanisms in Unicode.

The layouts are designed for ergonomy, typing speed and quality Unicode production. There are adaptations, all equivalent, for writers of Amazigh living or travelling in different countries. Previous tools concentrated on North Africa and France. There is also a bilingual alternative, closer in layout to the French AZERTY keyboard, for users preferring fewer changes to their customary layout or who frequently mix French and Amazigh.

Tifinagh and Latin layouts have been harmonised to encourage use of Tifinagh. This resulted in a Tifinagh-default version suitable for Morocco and an equivalent Latin-default version suitable for Algeria. From a Moroccan user's point of view the Tifinagh-default version extends Ircam's Tifinagh keyboard with international facilities (focusing primarily on other Amazigh language variants), rather like how the US international keyboard extends the US layout.

The layouts are intended for common use for Northern Amazigh in Latin and Tifinagh, and for common use typing Tifinagh for Tuareg and for transcribing Tuareg in Algerian Latin script. The West African Latin support is for Tuareg visitors to the north, and to help northern learners of Tuareg.

The keyboard layouts resulting from the project make it easy to build Tifinagh fonts independently of Amazigh language variants, opening up creative possibilities in typography. Fonts can use any coherent selection of Northern or Tuareg letters, perhaps suited to a particular region, and with any selection of the available typographical innovations. A set of Algerian fonts is under development to demonstrate this. Even if future practice moves away from the project's interpretation of Tifinagh in Unicode, the project will have enabled better understanding of Amazigh orthography and encouraged better typography.

By clarifying areas of Amazigh phonology and orthography, examining Latin and Tifinagh scripts side by side, and encouraging tool reuse across Amazigh variants, this work could contribute to future developments in Amazigh orthography and keyboard design. Examples of future work could be mobile phone keypad layouts like the Tifinagh keypad of the Sony Ericsson / Maroc Telecom J110i/J120i.

---

[4] Association for the Promotion of Tifinagh, Agadez, Niger

References

Ameur M. (2004). Les caractéristiques phoniques de l'Alphabet Tifinaghe-Ircam. In Ameur, M. and Boumalk, A., editors, *Proc. of conference 'Standardisation de l'amazighe' (Rabat, 8-9th December 2003)*, IRCAM, Rabat, p. 106.

Ameur M., Bouhjar A., Boukhris F., Boukous A., Boumalk A., Elmedlaoui M., Iazzi E. (2006). *Graphie et orthographe de l'amazighe*. IRCAM, Rabat.

Anderson P. (2010a). Evolution of the Tifinagh script in Unicode.

L2/10-113, Unicode Consortium.Tiffing

Anderson P. (2010b). Proposal to add two Tifinagh characters for vowels in Tuareg language variants. N3870 (L2/10-270), Unicode Consortium.

Andries P. (2004). Proposal to add the Tifinagh Script. N2739R, Unicode Consortium.

Bellil R. (2006). *Textes zénètes du Gourara*. CNRPAH, Algiers.

Bouhjar A. (2004). Le système graphique Tifinaghe-Ircam. In Ameur, M. and Boumalk, A., editors, *Proc. of conference 'Standardisation de l'amazighe' (Rabat, 8-9th December 2003)*, IRCAM, Rabat, pp. 51-54.

Bouridj N. and Nouh A. (2000). *Haqbaylit n Tipaza*. Éditions Tira.

Brugnatelli V. (2002). Tamazight et Unicode. La standardisation dans le domaine des ordinateurs. In Laceb M., editor, *Proc. of conference 'Tamazight face aux défis de la modernité' (Boumerdès, 15-17th July 2002)*, Algiers, pp. 215-227.

Delheure J. (1984). *Dictionnaire Mozabite-Français*. SELAF, Paris.

Elghamis R. (2004). *Guide de lecture et d'écriture en tifinagh vocalisées*. APT, Agadez, Niger.

Elghamis R. (2010). *Le tifinagh au Niger contemporain: Étude sur l'écriture indigène des Touaregs*. PhD thesis. University of Leiden.

Everson M. (1998). Encoding the Tifinagh script. Working Group Document N1757, Unicode Consortium.

Heath J. (2005). *A Grammar of Tamashek (Tuareg of Mali)*. Mouton de Gruyter.

Issouf, M. (2007). Les caractères *Tifinagh* dans Unicode. In *Proc. of conference 'Le libyco-berbère ou le Tifinagh' (Algiers, 21-22nd March* 2007*)*, HCA, Algiers, pp. 241-254.

Louali N. (1993). Les voyelles touarègues et l'alphabet tifinagh : évaluation de quelques propositions récentes. *Pholia, CRLS-Université Lumière Lyon 2*, vol. 8: 121-139.

Sudlow D. (2001). *The Tamasheq of North-East Burkina Faso*. Rüdiger Köppe Verlag, Köln.

Si tous les chemins mènent à Rome, ils ne se valent pas tous. Le problème d'accès lexical

Zock, Michael[1], Schwab, Didier[2]

[1]Laboratoire d'Informatique Fondamentale (LIF), CNRS & Aix-Marseille Université
michael.zock@lif.univ-mrs.fr

[2]Laboratoire d'Informatique de Grenoble, équipe GETALP, Université Grenoble 2
didier.schwab@imag.fr

Résumé

Tout le monde a déjà rencontré le problème suivant : on cherche un mot (ou le nom d'une personne) que l'on connaît, sans être en mesure d'y accéder à temps. Les travaux des psychologues ont montré que les personnes se trouvant dans cet état savent énormément de choses sur le mot recherché (sens, nombre de syllabes, etc.), et que les mots avec lequel ils le confondent lui ressemblent étrangement (lettre ou son initial, catégorie syntaxique, champ sémantique, etc.).

L'objectif de notre travail est de réaliser un programme tirant bénéfice de cet état de fait pour assister un locuteur ou rédacteur à (re)trouver le mot qu'il a sur le bout de la langue. À cette fin, nous prévoyons d'ajouter à un dictionnaire électronique existant un index d'association (collocations rencontrées dans un grand corpus). Autrement dit, nous proposons de construire un dictionnaire analogue à celui des êtres humains, qui, outre les informations conventionnelles (définition, forme écrite, informations grammaticales) contiendrait des liens (associations), permettant de naviguer entre les idées (concepts) et leurs expressions (mots). Un tel dictionnaire permettrait donc l'accès à l'information recherchée soit par la forme (lexicale : analyse), soit par le sens (concepts : production), soit par les deux.

1. Le mystère de la production verbale

Les êtres humains sont de vrais prodiges en matière de parole (production de langage). Non seulement ils sont capables de trouver rapidement les formes adéquates pour exprimer leurs idées sous forme de mots et de phrases,[1] mais, ils sont capables de se livrer à cet exercice pendant des heures, sans jamais se fatiguer. Pourtant, produire du langage, notamment à l'oral (discours spontané en temps réel) est une véritable gageure. Jugez-en vous même. Pour pouvoir s'exprimer à un débit normal un locuteur doit pouvoir localiser dans sa mémoire le mot exprimant sa pensée,[2] l'adapter morphologiquement, l'insérer au bon endroit de la phrase, tout en continuant à planifier l'idée suivante, et tout ceci en très peu de temps. Si jamais une de ces étapes tarde ou échoue, on assiste à des lapsus, bafouillages, sons de remplissage, ou, des pauses plus ou moins prononcées, pouvant aller jusqu'au silence total.

Vu le nombre de contraintes et le manque de temps, il est étonnant de voir le peu de fautes, notamment au niveau lexical (recouvrement du sens par des mots). En effet, si nos discours oraux sont truffés de fautes et d'imperfections de toutes sortes (hésitations, faux départs, pauses), ces dernières concernent rarement le niveau lexical : on ne se trompe sur les mots qu'environ une fois sur mille (Rossi et Peter-Defare, 1998). Comment est-ce possible ?

2. Explication possible, concernant une des tâches principales : l'accès lexical

La réponse à la question posée réside très vraisemblablement dans l'organisation de notre dictionnaire mental (Aitchinson, 2003) et dans l'efficacité des processus

---

[1]Il est tout à fait courant de produire spontanément un discours de 150-200 mots par minute, débit qu'on arrive à doubler en cas de besoin.

[2] Ce qui veut dire qu'on doit chercher dans un stock énorme (les chiffres avancés varient selon les auteurs entre 30 à 60 000 mots, voire plus) un élément particulier. C'est la fameuse aiguille dans une meule de foin. Il est clair que le nombre avancé est problématique pour diverses raisons : définition du terme « mot », connaissances actives/passives ; polysémie des mots, etc. Néanmoins, si ce chiffre vous paraît élevé, notez que le "Lexique anglais/français des sports olympiques", destiné aux journalistes couvrant les jeux de Sidney en l'an 2000, contenait déjà presque 14 000 mots, avec pas moins de 1000 entrées rien que pour les *sports aquatiques* (rubrique natation). Il n'y a aucun doute, la performance reste impressionnante, équivalant à la consultation d'un dictionnaire comme *Le Grand Robert* trois fois par seconde, et ceci pendant plusieurs heures.

de recherche misent au point lors de notre existence et lors de notre contact avec la langue. Vu cette efficacité remarquable, il paraît tout à fait indiqué, voire souhaitable de s'en inspirer et de considérer le cerveau,— sa structure et son fonctionnement— comme modèle, susceptible de nous aider à améliorer des béquilles cognitives que sont les dictionnaires électroniques pour assister des êtres humains à trouver le mot resté bloqué sur le bout de leur langue (Brown et Mc Neill, 1966). Car, si nous commettons peu de fautes, il nous arrive néanmoins de ne pas pouvoir localiser à temps un terme, et, dans ce cas, il est bien utile de pouvoir consulter une ressource externe, susceptible de nous révéler rapidement (en peu d'étapes) l'objet recherché. Avant de présenter cette solution (travail en cours), essayons de voir pourquoi nous échouons de temps en temps, n'arrivant pas à produire ce mot, en apparence à notre portée, mais restant finalement bloqué sur le bout de notre langue. Pour mieux comprendre ce qui se passe nous allons nous tourner un instant vers des travaux faits par des psychologues.

3. La production du langage vue par des psychologues

Pour comprendre le problème d'accès lexical, il faut le situer dans le cadre de sa tâche normale : la production de phrases.

Produire du langage consiste en gros à faire trois choses: concevoir un message, le traduire en langue, communiquer ce résultat sous forme graphique ou orale. C'est précisément ce qu'on retrouve dans la proposition de Garrett (1980, 1991) qui est à la base de tous les modèles proposés par des psychologues (Bock, 1995 ; Fromkin, 1993 ; Levelt 1989, 1993).[3] Il y aura donc un *conceptualiseur* (message), un *formulateur* (structure linguistique) et un *synthétiseur* de la parole (articulation). Même si l'ordre peut être bouleversé (rétroaction d'un niveau inférieur vers le niveau supérieur), l'ordre naturel est bien celui indiqué : on commence par les idées, pour terminer par la forme linguistique (sons, graphèmes). A noter, le passage des idées à la forme n'est pas direct, il est médité par la langue, notamment le choix lexical. C'est d'ailleurs surtout ce module intermédiaire qui a retenu l'attention de Garrett, car les traitements linguistiques laissent des traces (hésitations, erreurs). Ceci étant, il s'est donc appuyé sur une grande base de données d'erreurs pour construire son modèle.

---

[3]À noter, que le modèle utilisé en TAL est un peu différent. Il a été conçu par des linguistes-informaticiens (Reiter et Dale, 2000).

La tâche du **conceptualiseur** consiste à élaborer un *message* (conceptualisation) afin de réaliser un but ou une intention de communication. Cette structure ou forme de représentation est conceptuelle. C'est sur elle que s'effectueront les opérations linguistiques, précisant ainsi progressivement une structure qui à ce stade est sous spécifiée. [4]

Le **formulateur** prend en charge des aspects fonctionnels, positionnels et phonologiques des éléments utilisés pour communiquer le message.

Le niveau fonctionnel est responsable de l'*encodage grammatical*. C'est-à-dire, les concepts seront remplacés par des mots, ou plus précisément par des lemmes, auxquels on assigne le rôle qu'ils doivent jouer au sein de la phrase. Les lemmes ne sont pas encore des mots au sens classique du terme. Ils manquent d'informations, notamment phonologiques. Ce sont des représentations abstraites, contenant des informations sémantiques et syntaxiques : *catégorie lexicale* (nom, verbe, adjectif, etc.), *fonction syntaxique* (sujet, objet, etc.), *type de structure* dont le lemme peut faire partie (syntagme nominal, syntagme verbal, etc.), ainsi que certains *traits* ou *caractéristiques* spécifiques à la langue (ex. le genre). Ayant récupéré des lemmes auxquels on a assigné un rôle syntaxique on produit une *représentation fonctionnelle* de la phrase.

À l'étape suivante (*encodage phonologique*), on détermine la *représentation positionnelle*, c'est-à-dire, on récupère la forme phonologique, les caractéristiques segmentales et prosodiques des lemmes (qui, du coup deviennent des lexèmes) et on spécifie l'ordre des mots en les intégrant dans la structure spécifiée à l'étape précédente. C'est ici que seront insérés les morphèmes grammaticaux (ex. déterminants, flexions de nombre/genre/temps, prépositions, etc.) et que seront effectuées les opérations morphologiques. La structure issue de ce stade spécifie donc la position et l'articulation de l'ensemble des éléments de la phrase (mots

[4] Qu'il en soit ainsi est lié à des contraintes cognitives (Zock 1996). Des limites d'espace (mémoire de travail) et de temps (pression de production, manque de temps) font qu'on évite des engagements forts au début du processus. En effet, si l'on prenait très tôt des engagements forts on risquerait de s'enfermer dans des sens uniques, ne pouvant terminer une phrase, dont les éléments choisis au début se révèleraient incompatibles avec ceux de la fin. Autrement dit, on a intérêt de partir d'une structure sous-spécifiée, coquille relativement vide, qu'on enrichira ensuite au fur et à mesure en fonction des besoins.

pleins, mots grammaticaux), fléchis et accordés selon les règles de la grammaire. La tâche du niveau phonologique consiste non seulement à récupérer les phonèmes et à déterminer la prosodie de toute la séquence de mots mais également de traduire cette forme en un format (gestes phonatoires) susceptible d'être exécuté par l'articulateur.

L'**articulateur** doit transformer les symboles du module précédent en sons, afin d'évoquer chez l'auditeur des idées correspondantes à celles ayant donné naissance aux paroles du locuteur.

4. Les mots pièces toutes faites ou pièces détachées à assembler ?

Si pour un lexicographe les mots sont des entités, liant le sens et la forme, modèle qui remonte à Saussure, pour un psychologue ce sont des patterns distribués dans notre cerveau. Les modèles (Garrett, 1980, Levelt, 1989) et la décomposition de mots en lemmes et lexèmes est basée sur plusieurs types d'observations et d'expériences : l'analyse d'erreurs, le phénomène du 'mot sur le bout de la langue' (MBL) et la dénomination d'image.

Ayant analysé de nombreuses erreurs, Garrett (1991) a constaté qu'il y a deux grandes classes : celles touchant le *fond* ou le sens (erreurs de sélection : lion-tigre) et celles touchant la *forme* (erreurs d'assemblage : déchiffrer-défricher). Étant donné que les deux sont bien distinctes,[5] il en a déduit qu'il doit y avoir deux étapes (ou composantes) qui se succèdent dans le temps pour récupérer et déterminer la forme d'un mot : la première est responsable des aspects semantico-syntaxiques, le *lemme*, la seconde est chargée de récupérer les aspects phonologiques, la forme du mot (*lexème*).

Normalement les étapes de la détermination du sens, celle de la forme syntaxique et phonologique se suivent l'une après l'autre, mais il peut y avoir des dysfonctionnements, résultant en un blocage particulier, nommé 'mot sur le bout de la langue'. On connaît le sens, on connaît le mot et on en est conscient, pourtant, on

---

[5] Ainsi on observe des confusions du type sémantique '(orange/citron) et phonologiques (s'épanouir/s'évanouir', Laurence/Clémence) mais pour ainsi dire jamais, les deux à la fois. Ceci dit, deux mots peuvent avoir les deux types de liens (chat, rat), ce qui rend délicat l'analyse de la cause d'un éventuel dysfonctionnement. Était-il dû à une activation sémantique, phonologique ou les deux ?

n'arrive pas à récupérer dans un laps de temps raisonnable la forme correspondante. Des études sur le MBL ont montré que des locuteurs se trouvant dans cet état peuvent révéler bon nombre de propriétés concernant le mot recherché (sens, catégorie lexicale, genre grammatical, nombre de syllabes, etc.) sans pourtant être capable de le dénommer (Brown et McNeill, 1966). Autrement dit, ils ont accès au lemme, mais ils ne parviennent pas à récupérer le lexème, la forme phonologique correspondante.

Il y a une autre étude plaidant sur la distinction lemme/lexème. Dans une tâche consistant à demander aux sujets de nommer une image Schriefers et al. (1990) ont étudié les effets qu'un terme (selon qu'il est lié ou non au terme cible) et le moment de sa présentation peuvent avoir sur le résultat (accès lexical).

Les participants devaient dénommer des images (par exemple, RENARD, GOMME, RASOIR) tout en écoutant des mots distracteurs (lapin, pomme, blanco). Ces derniers leurs étaient présentés avant, pendant ou tout juste après la présentation de l'image. Selon la nature du terme distracteur (relation sémantique, phonologique, ou aucune relation avec le terme cible) et selon le moment de présentation (avant/pendant/après) on observait des effets différents. Un distracteur sémantique présenté avant le début de l'image produisait une interférence sémantique, tandis qu'un terme phonologiquement lié au mot cible avec un effet facilitateur si, et seulement s'il suivait l'image. On peut donc bel et bien conclure que les mots ne sont pas des entités holistiques, mais des structures qu'on précise progressivement. Il y a donc (au moins) deux niveaux ou deux types d'entités, des lemmes et des lexèmes, les deux étant sensibles chacun à des informations différentes, l'un aux informations sémantiques et l'autre aux informations phonologiques.

5. Améliorer la navigation dans les dictionnaires électroniques en s'inspirant du cerveau.

Comme nous avons pu voir, la représentation des mots dans des dictionnaires et dans le cerveau n'est pas la même. Ce dernier offre une série de caractéristiques, intéressantes à utiliser (haute connectivité, liens associatifs). Nous utiliserons donc la métaphore du *dictionnaire mental* (structure, processus, construction), pour montrer comment ce type de modèle pourrait nous guider à développer des aides à la navigation dans des dictionnaires électroniques. Ces derniers se prêtent, hélas, toujours assez mal à la production, dans la mesure où ils ne permettent pas l'affinement progressif d'une requête, et dans la mesure où l'on n'a toujours pas clarifié en quels termes communiquer le sens, point de départ en production.

Contrairement à la lecture on part du sens et non pas de la forme lexicale, celle-ci doit justement être trouvée.

5.1 Quelques caractéristiques concernant le dictionnaire mental

Une des questions qui se pose est de savoir à quoi ressemble (ne serait-ce que métaphoriquement) notre dictionnaire mental. Que savons-nous à l'heure actuelle ? Voici quelques traits caractéristiques :

- C'est un réseau dont les nœuds sont des mots [6] et les liens sont des associations liant les deux termes. L'idée remonte à Aristote,[7] et c'est une des bases de l'associationisme. Cette intuition a engendré un très grand nombre de travaux, dont WordNet est un des produits les plus aboutis (Miller, 1995).
- Le réseau est multidimensionnel et multi-couches. Les différents modules correspondant en gros aux modules classiques en linguistique informatique : *sens* (abstrait), *forme abstraite* (forme intermédiaire), *forme concrète* (son, graphème). Selon la tâche on part du 'sens' (production) ou de la 'forme' (analyse) pour arriver à leur correspondant : mots, concepts. L'information concernant un mot (sens, forme, son) est donc distribuée dans le réseau. Pour plus de détails, voir (Levelt, 1989 ; Levelt et al., 1999 ; Ferrand, 2002).
- La répartition des nœuds n'est pas uniforme. On remarque comme dans bien d'autres domaines (systèmes biologiques, réseaux sociaux) que les objets ne sont pas repartis de manière égale. Certains endroits du graphe ont une très

---

[6] Cette remarque doit être nuancée, non seulement parce que le terme 'mot' pose un certain nombre problèmes, mais aussi parce qu'on ne sait pas de façon claire si l'équivalent d'un lemme est stocké au niveau des nœuds, ou s'il y a seulement des liens allant vers d'autres nœuds, plus élémentaires (phonèmes, graphèmes), liens qui doivent être activés pour rendre accessible le lemme. Dans ce cas, celui-ci ne serait qu'une catégorie abstraite, donc bien différente de la forme concrète observée dans des dictionnaires habituels où la forme concrète d'un mot apparaît à côte de sa définition (sens).

[7] En effet, cette intuition se trouve déjà chez Aristote (« De memoria et reminiscentia »), puis chez des *philosophes* (Locke, Hume) et *physiologistes* anglais (James et Stuart Mills), des *psychologues* (Galton, 1880 ; Freud, 1901 ; Jung & Ricklin, 1906) et des *psycholinguistes* (Deese, 1965 ; Jenkins, 1970). Enfin, cette idée est sous-jacente à WORDNET (Miller,1990), aux travaux connexionistes (Stemberger, 1985 ; Dell, 1986), aux hypertextes et au web (Bush, 1945 ; Nelson, 1967). Pour des synthèses en psycholinguistique voir (Hörmann, 1972 ; chapitres 6-10), pour des références plus récentes voir (Spitzer, 1999).

grande densité, ils sont hautement peuplés (champs sémantiques : termes de parenté, couleurs, termes de calendrier, etc.), d'autres, en revanche, sont bien moins étoffés. Les graphes lexicaux ont donc des caractéristiques de « petit monde » (Schnettler, 2009) : *répartition inégale* (grappes, îlots), *chemins* relativement *courts* entre les différents éléments. Ceci a une portée non négligeable pour la navigation. Tout semble accessible via un petit nombre de pas (Motter et al. 2002). Ainsi peut-on utiliser un raccourci pour passer d'un domaine à un autre, sans transiter forcément via tous les nœuds intermédiaires ? (voir figure 2).

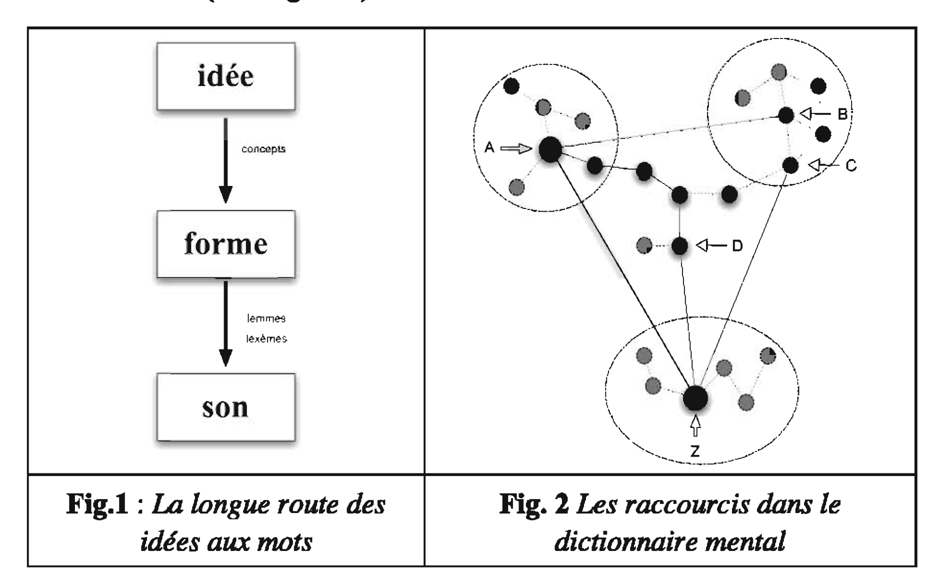


Fig.1 : *La longue route des idées aux mots*

Fig. 2 *Les raccourcis dans le dictionnaire mental*

5.2 Une carte mentale et une boussole lexicale pour assister l'orientation dans le réseau

Disposer d'un grand dictionnaire est de peu d'utilité si on ne peut pas accéder rapidement à l'information souhaitée. Or, c'est souvent le cas en production, dans la mesure où la plupart des dictionnaires ne permettent pas la consultation à partir du sens. Pourtant, c'est la situation la plus fréquente pour un rédacteur. Pour combler cette lacune, nous envisageons d'ajouter à un dictionnaire électronique existant un index basé sur les notions d'*association* et des *primitives* de *sens*. À la différence des *dictionnaires traditionnels,* où toute l'information concernant un mot est stockée directement avec cette entrée lexicale, les informations relatives aux mots sont distribuées dans le cas du *dictionnaire mental* (cerveau). C'est d'ailleurs pour cette raison qu'on a parfois tant de mal à les rassembler, en temps voulu, pour

aboutir à la forme finale de ce puzzle lexical. Comme des expériences portant sur le phénomène du « mot sur le bout sur la langue » l'ont bien montré, même en cas d'échec, l'utilisateur sait pratiquement toujours quelque chose concernant le mot convoité : sens, origine, mots liés par association, etc. Et c'est de cette information, mot disponible à cet instant, dont nous allons nous servir comme point de départ, pour entrer dans un réseau lexical, — le fait que les idées, ou les mots les exprimant s'évoquent réciproquement prouve bien que le dictionnaire mental est un réseau pour avancer, petit à petit, vers le mot recherché. Pour l'aider à s'orienter nous allons donc fournir certains outils de navigation, comme une carte (réseau lexical, balisé en terme de liens) et une boussole (index), et nous proposons des ponts de taille variable, permettant d'aller du *mot source* vers le *mot cible*, mot recherché.

Pour trouver le mot sur bout de la langue il nous faut donc une *carte mentale* et une *boussole lexicale* (Zock et al., 2010). On notera, que s'il y a des méthodes pour calculer le chemin le plus court dans un réseau, celles-ci sont de peu d'utilité. Car si le système connaît le *point de départ*,— il lui est donné via l'entrée (requête faite par l'utilisateur),— il ne connaît pas le *point d'arrivée*. En revanche, le système peut nous aider à le trouver, car, avec chaque information reçue, il peut nous proposer une série de candidats (termes directement associés, termes se trouvant à une distance de 1) susceptibles de contenir le mot cible ou un terme permettant de s'en approcher. Contrairement au système, l'utilisateur (locuteur) saura reconnaître le bon terme lorsqu'il le voit (voir les études sur le « bout de la langue, Brown et Mc Neill, 1966). Autrement dit, même si ce n'est pas le terme exact, mais seulement un terme plus ou moins directement lié, l'utilisateur saura alors lequel d'entre eux pointe dans la bonne direction, ou lequel est le plus proche du mot recherché. Supposons qu'on cherche le mot 'infirmière', en donnant en entrée le mot 'hôpital'. L'utilisateur saura alors qu'aucun des termes suivants 'asile, hospice, clinique, sanatorium' n'est la bonne solution. En revanche, la liste décrivant des *employés de l'hôpital* (médecin, anesthésiste,…) est susceptible de contenir le terme recherché : infirmière. Il en est ainsi pour 'infirmerie'. Trouver le mot cible à partir d'un mot source quelconque, disponible à ce moment, suppose une carte mentale, décrivant les différents types de lien entre les objets du monde (connaissance encyclopédique). Comment construire une telle carte et comment s'y orienter grâce à une boussole lexicale sera l'objet de notre exposé.

6. Conclusions

Un *dictionnaire* est un composant fondamental de tout système de traitement de la langue, sa qualité dépendant des *informations* stockées et des moyens offerts pour y accéder. Or, les stratégies d'accès dépendront de nombreux facteurs : *connaissances* disponibles lors de la consultation (sens, mots réliés au mot cible, etc.), *tâche* (analyse vs. production) et *nature* du système de traitement (humain, machine). En *analyse,* on part des mots pour chercher le sens, tandis qu'en *production,* on part des concepts pour trouver les mots correspondants. Et si une machine trouve généralement l'information stockée (accès lexical en génération par une approche TAL (Stede, 1999), un être humain ne saura pas forcément faire autant : trop nombreuses et trop différentes sont les informations à traiter en si peu de temps (discours spontané).

Il y a au moins trois types de dictionnaires : les dictionnaires papiers, les dictionnaires électroniques, et les dictionnaires mentaux (cerveau). Ces derniers offrent certaines caractéristiques particulièrement intéressantes pour la production. La multiplicité de points de vue pour organiser le lexique (indexation selon différents points de vue ou de niveaux : sens, forme, son) offre du coup une souplesse d'accès inégalée.

Contrairement aux dictionnaires organisés de manière rigide par ordre alphabétique ou par inclusion (ordre hiérarchique), les dictionnaires mentaux sont des réseaux dont les termes sont hautement connectés, du coup tout peut être accédé à partir de n'importe quel point du réseau, et les liens assurant cette connexion sont les rôles que les termes jouent dans la vie réelle. C'est précisément cela qui nous permet d'accéder à l'information recherchée.

De ce fait le dictionnaire mental constitue un excellent modèle en termes de stockage et d'accès à l'information. Si les dictionnaires traditionnels sont passifs et assez limités en termes d'accès, les dictionnaires électroniques ont un potentiel considérable, susceptibles de présenter rapidement et sous des formes diverses l'information recherchée. Les idées présentées ici sont une première tentative allant dans ce sens, mais il est clair, que beaucoup de travail reste encore à faire, notamment au niveau des liens (associations).

Références

Aitchinson, J. (2003): *Words in the Mind: an Introduction to the Mental Lexicon.* Oxford, Blackwell.

Aristote (350 avant JC) *De memoria et reminiscentia.* In Parva Naturalia, Vrin

Bock, J.K. (1995). *Sentence production: From mind to mouth.* In J. L. Miller & P.D. Eimas (Ed.), Handbook of perception and cognition. Vol. 11: Speech, language and communication. Orlando, FL: Academic Press.

Brown, R & Mc Neill, D. (1966): *The tip of the tongue phenomenon.* Journal of Verbal Learning and Verbal Behavior, 5, 325-337.

Bush, V (1945) *As we may think.* The Atlantic Monthly; Volume 176, No. 1; pages 101-108

Deese J. (1965) *The structure of associations in language and thought.* Baltimore

Dell, G. S., Chang, F., and Griffin, Z. M. (1999), *Connectionist Models of Language Production: Lexical Access and Grammatical Encoding*, Cognitive Science, 23/4, pp. 517-542.

Ferrand, L. (2002) : *Modèles de la production de la parole.* In M. Fayol (Ed.) Production du langage. Traité des Sciences Cognitives. Paris : Hermès, 27-44.

Freud, S. (1901) *Psychopathologie de la vie quotidienne.* Paris : Payot, 1997

Fromkin, V. (1993). *Speech Production.* In Psycholinguistics. J. Berko Gleason & N. Bernstein Ratner, *Eds*. Fort Worth, TX: Harcourt, Brace, Jovanovich.

Galton, F. (1880). *Psychometric experiments.* Brain, 2, 149-162.

Garrett, M. (1991). *Sentence processing.* In D, Osherson and H. Lasnik (Eds.), Language: An invitation to cognitive science, Cambridge, Mass.: The MIT Press.

Garrett. M. F. (1980). *Levels of processing in sentence production.* In B. Butterworth (Ed.), Language production (pp. 177-220). London: Academic Press

Hörmann H. (1972) *Introduction à la psycholinguistique.* Paris, Larousse

Jenkins, J.J. (1970). *The 1952 Minnesota word association norms.* In: L. Postman, G. Keppel (eds.): Norms of Word Association. New York: Academic Press, 1-38.

Jung, C.G., Ricklin, F. (1906). *Experimentelle Untersuchungen über Assoziationen Gesunder.* In: C.G. Jung (ed.): Diagnostische Assoziationsstudien. Leipzig: Barth, 7-145.

Levelt W., Roelofs A. & Meyer, A. (1999). *A theory of lexical access in speech production.* Behavioral and Brain Sciences, 22, 1-75.

Levelt, W. (1993) *The architecture of normal spoken language use*. In Blanken G., J. Dittmann, H. Grimm, J. Marshall & C. Wallesch (eds.) Linguistic Disorders and pathologies. W. de Gruyter, Berlin, New York

Levelt, W.J.M. (1989): *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Lexique anglais/français des sports olympiques: jeux d'été, *Insep Publications*, Paris (2000).

Meyer A.S & Bock K. *The tip-of-the-tongue phenomenon: Blocking or partial activation*? Memory & Cognition. 1992;20:715–726.

Miller G. (1995). *WordNet: A Lexical Database for English.* Communications of the ACM Vol. 38, No. 11: 39-41

Motter, A. E., A. P. S. de Moura, Y.-C. Lai, and P. Dasgupta. (2002). *Topology of the conceptual network of language*. Physical Review E, 65(6).

Nelson, T. (1967) Xanadu Projet hypertextuel, http://xanadu.com/

Palermo, D., Jenkins, J. (1964). *Word Association Norms*. Minneapolis, MN: University of Minnesota Press.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge: Cambridge University Press.

Rossi M., & Peter-Defare, E. (1998). *Les lapsus, ou comment notre fourche a langué*. Paris : Presses Universitaires de France.

Rossi, M. (2001). *Les lapsus et la production de la parole*. Psychologie Française, n° 46, pp. 27-41.

Schnettler, S. (2009). *A structured overview of 50 years of small-world research*. Social Networks, Vol. 31, No. 3., pp. 165-178.

Schriefers H., Meyer A. S., Levelt W. J. M. (1990) *Exploring the time course of lexical access in production : Picture-word interference studies*, Journal of Memory and Language, 29, 86-102.

Spitzer, M. (1999). *The mind within the net : models of learning, thinking and acting*. A Bradford book. MIT Press, Cambridge

Stede, M. (1999). Lexical semantics and knowledge representation in multilingual text generation Boston/Dordrecht/London: Kluwer Academic Publishers, 1999

Stemberger, J. P. (1985) *An interactive activation model of language production*. In A. W. Ellis [ed] Progress in the Psychology of Language, Vol. 1, 143-186. Erlbaum.

Zock M, Ferret, O. and Schwab D. (2010). *Deliberate word access : an intuition, a roadmap and some preliminary empirical results*. International Journal of Speech Technology, 13(4): 107-117

Zock, M. (1996). *The Power of Words in Message Planning*, 16th International Conference on Computational Linguistics, Copenhagen, pp. 990-995

Workshop on Computational Approaches to Semitic Languages, pages 1–8, Ann Arbor.

(Xu J., Croft B. W, 1998) Xu J., Croft B. W. Corpus-based stemming using co-occurrence of word variants. ACM Transactions on Information Systems, 1998.

المراجع

(Buckwalter, T., 2004) Buckwalter, Tim . Buckwalter Arabic Morphological Analyzer Version 2.0. 2004. Linguistic Data Consortium (LDC) catalog number LDC2004L02, ISBN 1-58563-324-0.

(Chen, A., Gey, F., 2002) Chen, A., Gey, F. "Building an Arabic stemmer for information retrieval".TREC-11 conference, 2002.

(Darwish, K, 2002) Darwish, K. "Building a Shallow Arabic Morphological Analyzer in One Day". In The ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, USA, 2002.

(De Roeck, A. N., Al-Fares, W, 2000) De Roeck, A. N., Al-Fares, W. "A morphologically sensitive clustering algorithm for identifying Arabic roots". In Proceedings ACL-2000. Hong Kong, 2000.

(Diab, M. W et al, 2007) Diab, M., Ghoneim, M. and Habash N. Arabic Diacritization in the Context of Statistical Machine Translation, In Proceedings of the Machine Translation Summit (MT-Summit), Copenhagen, Denmark, 2007.

(El Sadany, T.A, Hashish, M.A., (1989El Sadany, T.A & Hashish, M.A.: 1989, An Arabic Morphological System. IBM System Jornal.Vol 28, NO4.

(Kenneth R. Beesley, 2001) Kenneth R. Beesley: 2001, Finite-State Morphological Analysis and Generation of Arabic at Xerox. Xerox Research Centre Europe. MEYLAN, France.

(Khoja, S., Garside, 1999) Khoja, S., Garside, R "Stemming Arabic text", Computing Department, Lancaster University, Lancaster, 1999. ww.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps.

(Larkey, L. S. et al, 2002) Larkey, L. S., Ballesteros, L. & Connel, M. E. "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis", in Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002.

(Mayfield J., McNamee P., 2003) Mayfield J., McNamee P. "Single N-gram Stemming", SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003.

(Soudi, A., 2005) Soudi, A: 2005, Memory-based Morphological Analysis Generation and Part-of-speech Tagging of Arabic. Proceedings of the ACL

بدلا من" مرما" الواردة في كلمة سطحية من قبيل "مرماهم". أما في حالة وجود ياء بعدها ألف بمفرده أو متلو بحروف أخرى (أقصاها ثلاثة)، يحذف ما بعد الياء وتقلب ألفا مقصورة، وينصب البحث عن هذه الكلمة في قاموس الكلمات (وتحديدا في الصنف الفرعي1 المتعلق بالمقصور). فإذا ما وجدت، تكون عملية القلب صحيحة ويتم التحقق من أن ما بعد الياء هو بالفعل لاحقة استنادا إلى قاموس اللواحق، وإذا لم يكن موجودا، تلغى عملية القلب مباشرة.

2) بالنسبة للمنقوص (كمحام ومحامون)، يتم البحث للوهلة الأولى، وكالعادة، في قاموس الأسماء الذي لا يتوفر على ما يبدو على كلمات من هذا القبيل. عندئذ تتم إضافة "ي" إلى هذه الكلمة بعد إزالة اللاحقة (إذا كانت موجودة) وقبل البحث عنه في الصنف الفرعي2 الخاص بالمنقوص. وإذا ما تم العثور على الكلمة في قاموس الأسماء منذ الوهلة الأولى، فلن تضاف الياء إلى الجذع مطلقا.

3) تبديل الشكل "ؤ" والشكل "ئـ" بالشكل الأصلي "ء" والبحث عن الكلمة في الصنف الفرعي3 المتعلق بالمهموز إذا كانا متلوين بحروف موجودة في قاموس اللواحق، وإلا فهي أصلية ولا ينبغي تبديلها.

4) في حالة الكلمات التي تبدأ بـ "ال"، تحذف اللام الأولى ويتم البحث عن البقية في قاموس الأسماء، فإذا وجد فهذا يعني أن هذه اللام يمكن أن تكون "ل" بمفردها أو "لل" (التي تفكك استنادا إلى قاموس السوابق على النحو الآتي: ل/ال)، وأن اللام الثانية هي أصلية. أما إذا تعذر وجودها، تحذف لل ويتم التحقق من وجود الباقي في قاموس الأسماء.

5) كل كلمة تنتهي بالتاء المبسوطة "ت" بعد تجريدها من جميع اللواصق يتم البحث عنها في بداية الأمر في قاموس الأسماء، وسواء وجدت أو لم توجد تحول "ت" إلى "ة"، مما يتطلب القيام ببحث جديد عن هذه الكلمة الجديدة في قاموس الأسماء. وفي حالة جمع المؤنث السالم الذي ينصب على الأسماء التي تنتمي للطبقتين الأولى والثانية، يتم حذف لاحقة العدد المؤنث من الكلمة وإضافة تاء مربوطة لما تبقى منها قبل البحث عنها في قاموس الأسماء.

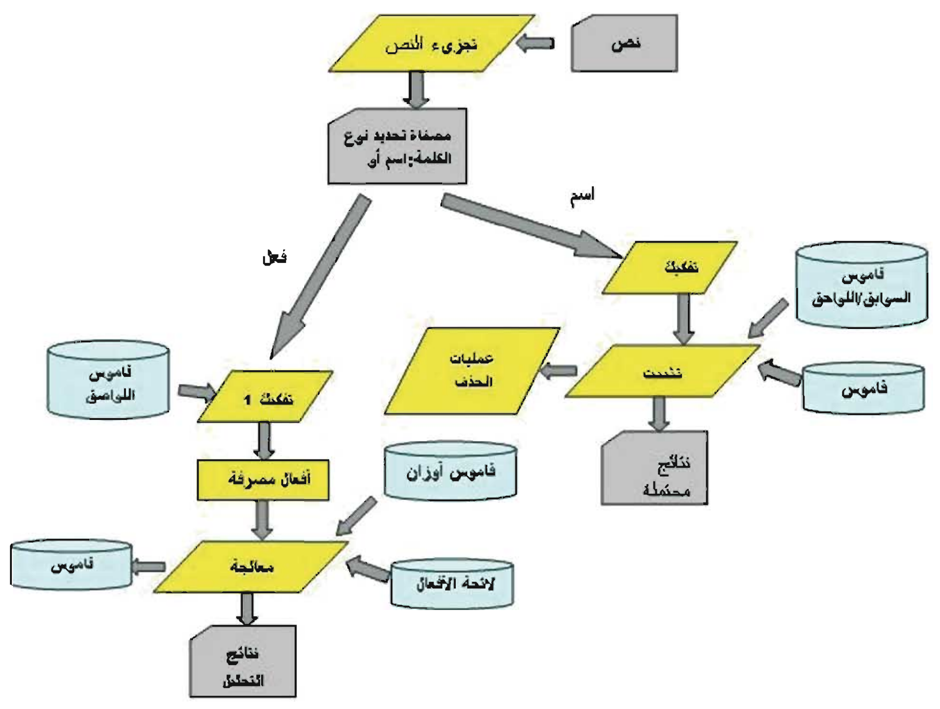
6) تكون عملية التفكيك مشروعة فقط إذا كان ما تبقى من الكلمة بعد إزالة اللواصق لا يقل عن حرفين.

7) أية كلمة تنتهي بالياء ينطبق عليها ما يلي: إذا كانت موجودة في الصنف الفرعي2، فإن هذه الياء أصلية ولا يمكن أن تكون لاحقة.

خاتمة

أثناء المعالجة الآلية للغة العربية لا بد من مراعاة الدقة وتوخي الحذر لكثرة تمظهراتها الناتجة بالدرجة الأولى عن كونها غير مرفقة بالحركات القصيرة. وقد قدمنا في هذه الورقة ما نعتقد أنه أساس لساني متين مكننا من التوصل إلى تنظيم معقول للعناصر المدرجة في قاعدة المعطيات اليدوية لبلوغ الكفاية والشمولية والإحاطة بكل الاحتمالات الممكنة. فكان النظام المقترح أدنويا من ناحيتين: أ) اعتماد قاموس مصنف للأسماء لا يتضمن الكلمات السطحية التي يمكن اشتقاقها بإضافة لواصق إليها، ب) اعتماد النظام على إجراءين تكراريين مرتبطين بعدد محدد من عمليات الحذف والإبدال، هما: التفكيك والعودة للقواميس (قاموس الأسماء على وجه الخصوص).

يتم البحث عن كل كلمة سطحية في قاموس الأسماء أولا قبل أن تفكك بالنظر إلى أحد القواميس المعتمدة، ثم تفكك على أساس قاموس السوابق (لنحصل على: سابقة/كلمة محتملة)، وتتكرر هذه العملية حتى تشمل جميع السوابق الواردة، وفي كل مرة يتم البحث من جديد عن ما تبقى من الكلمة في قاموس الأسماء. بعد ذلك تنطبق قاعدة التفكيك مرة أخرى، ولكن على أساس قاموس اللواحق هذه المرة (لنحصل على: كلمة محتملة/لاحقة) فيتم اللجوء إلى قاموس الأسماء للمرة الثالثة، وبشكل متكرر، للبحث عن ما تبقى بعد إزالة اللواحق بكيفية تدريجية على الشاكلة المشار إليها قبل قليل. وعند القيام بعملية التفكيك الثالثة والأخيرة سيتجرد الجذع من كل اللواصق العالقة به باعتماد القواميس الثلاثة (لنحصل على: سابقة/كلمة محتملة/لاحقة). في هذه المرحلة يتم احتساب كل التأليفات الممكنة بين العناصر الثلاثة المكونة للكلمة السطحية بالعودة المتكررة لقاموس الأسماء



للتحقق من وجودها فيه.

خطاطة نظام التحليل

عمليات القلب والحذف

لمعالجة جملة من التغيرات التي تطرأ على الكلمات السطحية نقترح هذه العمليات التي من شأنها أن تسعفنا في الإلمام بجل السيناريوهات الممكنة:

1) في حالة وجود ا (الألف) بعدها ثلاثة حروف على الأكثر تقلب "ى" (الألف المقصورة) شريطة أن يكون ما بعدها موجودا في قاموس اللواحق، مما يعني أن البحث سينصب على الكلمة "مرمى"

نموذج من قاموس الكلمات المعتمد

الجذر	الكلمة	الكلمة المشكولة	الطبقة	الصنف	العدد	الجنس
ك ت ب						
	كتب	كَتَب	1	2	1	1
	كتب	كُتُب	1	4	2	2
	كتاب	كِتَاب	1	1	1	1
	كتاب	كُتَّاب	1	4	2	1
	كتاب	كُتَّاب	1	1	1	1
	كتابة	كِتَابَة	1	2	1	2
	كتيب	كُتَيِّب	1	1	1	1
	مكتب	مَكْتَب	1	1	1	1
	مكتبة	مَكْتَبَة	1	1	1	2
	اكتتاب	اكْتِتَاب	1	2	1	1
	..	..	..	..	..	..
ك ت ح						
	كتح	كَتَح	1	2	1	1

3.2.2 كيفية اشتغال قالب تحليل الأسماء

بعد حصوله من المجزء على كلمات سطحية (أسماء) حاملة لكل العناصر الملصقة بها حيث يتعامل معها الواحدة تلو الأخرى، إذ يفككها إلى عناصرها الأساسية ويثبت مكوناتها المفككة بالاعتماد على القواميس الثلاثة: قاموس الأسماء، وقاموس السوابق، وقاموس اللواحق، وعلى عمليات الحذف والقلب التي تلعب دورا مركزيا في التعرف على الكلمة وتسويغ النتائج المحتملة. ويقدم المحلل في نهاية المطاف النتائج المحتملة التي يتم العثور عليها. أما إذا تعذر عليه ذلك، فذلك يعني أن هذه الكلمة السطحية غير صحيحة، أو أنها غير موجودة في قاعدة المعطيات اللغوية (وفي هذه الحالة ينبغي ضمها إليها). وعليه، فالهيكلة الداخلية لهذا المحلل واردة في الخطاطة أسفله.

2.2.2 الموارد اللسانسة لقاعدة المعطيات اليدوية

أ) القواميس

هناك ثلاثة قواميس يتضمن كل واحد منها عنصرا من العناصر الثلاثة المكونة للكلمة السطحية (أو الكلمة الدخل): السوابق واللواحق والجذوع. فقاموسا السوابق واللواحق يحتويان على كل اللواصق الاسمية البسيطة والمركبة الموجود في اللغة العربية، وهي مشكولة وغير مشكولة، مفككة وغير مفككة ومصحوبة بمصفوفة من المعلومات المتعلقة بها. ويقوم قاموس الأسماء على تقسيم ثلاثي الطبقات تندرج ضمنه كل المركبات الاسمية التي يمكن العثور عليها في اللغة العربية، وهو مرتب ألفبائيا على أساس الجذر (الصامتي) الذي يحتضن قائمة من الكلمات المصصفة أسفله حسب الترتيب الألفبائي أيضا. هذه الكلمات مشكولة وغير مشكولة ومقرونة هي الأخرى بالمعلومات المتعلقة بالطبقة والصنف والصنف الفرعي، فضلا عن سمتي الجنس والعدد.

وضمن الأصناف الاسمية المعتمدة نجد أسماء ذات طبيعة خاصة لكونها تظهر في أشكال مختلفة، كما أشرنا إلى ذلك من قبل، يتعلق الأمر هنا بالمهموز والمنقوص والمقصور والأسماء الخمسة التي تشكل ما نسميه بالأصناف الفرعية.[3]

الجنس والعدد

يوسم الجنس المذكر على مستوى القاعدة بالنسبة لجميع الطبقات باستثناء الحالات التي تكون فيها الكلمة السطحية مصرفة في المثنى المذكر أو جمع المذكر السالم، في حين أن الجنس المؤنث يوسم على مستوى القاعدة بالنسبة للطبقة الأولى وعلى مستوى اللاحقة بالنسبة للطبقتين الثانية والثالثة في المفرد والمثنى المؤنث وجمع المؤنث السالم.[4] وفي ما يخص العدد نشير إلى أن جميع الكلمات موسومة بالعدد المفرد على مستوى القاعدة، ما عدا تلك المدرجة في الصنف الرابع، يعني جموع التكسير التي تكون موسومة على مستوى القاعدة دائما على أنها جمع. وعندما تكون الكلمة السطحية مصرفة في المثنى والجمع السالم بنوعيهما، تستخلص سمة العدد المناسبة من نفس السمة الموجودة في اللاحقة (وهي سمة متضمنة في قاموس اللواحق).[5]

لدينا، إذن، ثلاث طبقات اسمية، وضمن كل طبقة هناك أصناف وأصناف فرعية موزعة على النحو الآتي: بالنسبة للصنف: 1= اسم الذات، 2= اسم الحدث، 3= اسم العلم، 4= جمع التكسير، 5= اسم الفاعل، 6= اسم المفعول، 7= صيغة المبالغة. بالنسبة للصنف الفرعي: 1= المقصور، 2= المنقوص، 3= المهموز. بالنسبة للجنس: 1= المذكر 2= المؤنث. بالنسبة للعدد: 1= المفرد، 2= جمع التكسير.

---

[3] نظرا لقلة الأسماء الخمسة في اللغة العربية ستزود قاموس الأسماء بكل تمظهراتها الممكنة.

[4] فكلمة من قبيل *مدرسة* أو *دار* موسومة في القاعدة على أنها مؤنث، بيد أن سمة الجنس المؤنث في كلمة نحو معلمة أو شجرة أو معلمتان أو شجرات مستخلصة من نفس السمة في اللاحقة الواردة.

[5] الهدف هو تقديم أكبر عدد ممكن من المعلومات حول عناصر الكلمة الدخل المفككة، والاستغلال المستقبلي لهذه المعلومات في إنجاز محللات تركيبية أو دلالية أو في وضع مترجمات آلية الخ.

وطبيعة اللواصق المسموح لها بالانضمام إليها. أما على المستوى المنهجي للتعامل مع النمط المقولي الاسمي، نشدد على ضرورة الانتباه لأمرين أساسيين، أحدهما هو أن النصوص العربية، التي من المفروض أن تشكل مجال عمل المحلل الصرفي، هي نصوص غير مشكولة، يلتبس فيها الاسم بالفعل وتتداخل فيها الأصناف الاسمية بعضها ببعض. والأمر الآخر هو أن عددا لا يستهان به من العناصر المدرجة ضمن لائحتي السوابق واللواحق تتحقق كحروف أصلية في بعض الأنماط المقولية الأخرى.

أصناف الأسماء التي اعتمدناها في نظام التحليل محصورة فيما يلي: اسم الجنس، واسم العلم، واسم الآلة، واسم المكان، واسم الزمان، والتصغير، واسم الحدث، واسم الفاعل، واسم المفعول، وصيغة المبالغة.

الأصناف الستة الأولى هي أسماء ذوات باستثناء اسم الزمان[2]، وهي تتقاسم الخصائص الآتية: أ) تقبل لاحقتي المثنى وجمع المؤنث السالم، ب) تجمع جمع تكسير فقط، ج) التاء المربوطة ليست لاحقة، ولا تنفصل عن الكلمة، ولا تعبر عن التأنيث بالضرورة، د) تختفي التاء المربوطة من الكلمة عند ما تتلوها لاحقة أو عندما تكون مجموعة جمع مؤنث سالم، ه) تقبل النسبة. وفي مقابل ذلك، يبدي اسم العلم سلوكا مخالفا من حيث أنه غالبا ما يرد في صورة صرفية واحدة قلما تتغير، لكنه لا يخرج عن إطار الخصائص السابقة الذكر شأنه في ذلك شأن اسم الحدث الذي لا يدل على الذوات، والذي لا يقبل أحيانا أن يثنى أو يجمع. وتختلف الأصناف المذكورة عن الثلاثة الأخيرة في كونها تملك الخصائص ذاتها: أ) التاء المربوطة لاحقة للتأنيث دائما، ب) النسبة غير ممكنة، ج) جمع المذكر السالم ممكن.

بناءا على ما سبق، قمنا بتجميع بضعة أصناف اسمية في طبقات كلما تشابهت في قبولها لنفس العدد من اللواحق، فكان أساس هذا التقسيم مستمدا من كون الأصناف الاسمية الموجودة في اللغة العربية تنتمي بالضرورة لإحدى الطبقات الثلاث الرئيسية الآتية:

- ✓ الطبقة الأولى، وهي التي لا تقبل لاحقة التأنيث *"ة"* ولاحقة الجمع المذكر السالم *"ون/ين"*، تتضمن الأصناف الاسمية التالية (وهي تشكل نسبة كبيرة من قاموس الأسماء): اسم الذات/الحدث/العلم/جمع التكسير.
- ✓ الطبقة الثانية، وهي التي لا تقبل لاحقة الجمع المذكر السالم *"ون/ين"*، ولاحقة المثنى *"ان/ين"*، تتضمن اسم النوع فقط، وهو إما اسم ذات أو حدث.
- ✓ الطبقة الثالثة، وهي الطبقة التي لا تقبل لاحقة النسبة، تتضمن الأصناف الاسمية التالية: اسم الفاعل/المفعول/صيغة المبالغة.

لقد أدرجنا جمع التكسير إلى جانب اسم الذات والحدث والعلم في الطبقة الأولى لكونه لا يخضع لقاعدة مطردة (في حالة الثلاثي)، ولكونه يتقيد بخصائص الطبقة التي ينتمي إليها. هناك بعض الأصناف الاسمية التي وردت في طبقتين، ونخص بالذكر هنا اسمي الذات والحدث الواردين في الطبقتين الأولى والثانية، ويرجع ذلك إلى أن بعض الأسماء المدرجة ضمن هذين الصنفين تقبل التاء المربوطة التي تكون صنيفة الوحدة، وهي لاحقة كما في قولنا: شجر-شجرة، رقص-رقصة. هذه الأسماء ينبغي أن تجمع في طبقة خاصة لا تتضمن اسم الوحدة.

---

[2] لقد أدرجنا اسم الزمان تجاوزا ضمن الصنف الخاص باسم الذات، على أن نقوم بفصله عنه مستقبلا.

الفعل التركيبي والدلالي كالواو والفاء ونحوهما بالنسبة للسوابق، وضمير المفعول كالكاف والهاء وغيرهما بالنسبة للواحق. وثانيهما يتمثل في السمات الصرف-تركيبية الواردة في اللائحة (ج)، وخصوصا ضمير الفاعل.

2.1.2 التحليل والتفكيك

عند تجزيء النصوص العربية إلى وحداتها السطحية وعزلها عن بعضها البعض، يقوم هذا الجزء من المحلل الصرفي بالتعامل مع الأفعال فقط دون غيرها من الأنماط المقولية الأخرى. ويستهل عملية التحليل بإزالة شكل الكلمات السطحية إذا كانت مشكولة، ثم يزيل الزمرة الأولى من اللواصق اعتمادا على لائحة اللواصق السابقة الذكر (أي الواو والفاء وضمير المفعول ونحوها). في هذه المرحلة من التحليل، إذن، يتم التخلص من جميع اللواصق باستثناء السمات الصرف-تركيبية المتمثلة أساسا في ضمير الفاعل (أي السمات التطابقية) ولاصقة الزمن (وتحديدا سمتي الوجه والبناء).

الشكل السطحي المحصل عليه بعد القيام بهذه العمليات يخضع لسيرورات إضافية تضمن بلوغ التحليل لكل النتائج المقبولة في الصرف العربي. لشرح هذه السيرورات نورد المثال الآتي:

وهكذا يقدم المحلل جميع النتائج الممكنة ذات الصلة بغياب الحركات في اللغة العربية، لأن خذا الغياب هو الذي يتيح تعدد الإمكانات الواردة خصوصاً وأن الحركات تلعب دورا رئيسيا في تصريف الفعل العربي.

وبعد الانتهاء من هذه المرحلة الأولى من التحليل، تنصب العمليات المتبقية على إزالة النمط الثاني من اللواصق (أي اللواصق الصرف-تركيبية) للحصول على الجذع ومن خلاله على الجذر، قبل تقديم الصيغة النهائية للنتائج المبلوغة.

3.1.2 نتائج التحليل

لا يقتصر هذا المحلل على تقديم النتائج الممكنة فحسب، وإنما يتجاوز ذلك إلى تقديم عدة معلومات تتعلق بمختلف اللواصق التي لحقت الجذع الفعلي. إنه يفرز السوابق عن اللواحق، ويحدد زمن الفعل وبنائه وإعرابه وعدده وجنسه وشخصه ووزنه الصرفي.

2.2 قالب معالجة الأسماء

للأسماء العربية خصوصيات كثيرة يتعلق البعض منها بخصائص اللغة العربية في حد ذاتها، بينما يتعلق البعض الآخر بالطبيعة المركبية للمقولات الاسمية.

1.2.2 التصنيف الاسمي المعتمد: الموارد والأسس

تتطلب الأسماء معالجة خاصة من شأنها المساهمة في بلورة تصور جديد يضبطها ويخول حوسبتها وترجمة خصائصها آليا. فعلى المستوى التصنيفي للأسماء ينبغي الانطلاق من عملية تصنيف الأسماء إلى أصناف تتقاطع في خصائصها العامة والبارزة (أي الشكلية)، وتتعارض في عدد

الأفعال	المقولة 1	المقولة 2	المقولة 3
حَفَزَ	1		
حَفَلَ	1	2	
حَفَدَ	1		
حَفَرَ	1		
. .			

قاموس الأفعال

	الزمن	البناء	الوجه	الشخص	الجنس	العدد
1	الماضي	المعلوم	---	1	محايد	مفرد
2	الماضي	المعلوم	---	1	محايد	مثنى
3	الماضي	المعلوم	---	1	محايد	جمع
. .						

لائحة السمات الصرف-تركيبية

ب) قاموس الأوزان، ويشتمل على الأوزان السطحية التي تساير البنية الشكلية للفعل المصرف، والأوزان الصرفية المعروفة. فالوزن السطحي، مثلا، لفعل من قبيل "أكل" هو "أعل"، و"قال" هو "فال"، و"دنا" هو "فعا"، و"جرى" هو "فعى" الخ، بينما يكون الوزن الصرفي لكل هذه الأفعال واحدا لا أكثر وهو "فعَل". إن مختلف الأوزان المتضمنة في هذا القاموس مولدة باعتماد مصرف آلي، وهي مصرفة آليا في جميع الأزمنة (الماضي، والمضارع، والأمر)، وفي البناء للمعلوم والبناء للمجهول، وفي المضارع المرفوع والمنصوب والمجزوم، ومع جميع ضمائر الشخص المستعملة في اللغة العربية.

ف ع ل	فَعَلَ	1	فَعَلَ	فَعَلَ	1	يَجْلِسُ	جَلَسَ	1
ف ع ل	فَعَلَ	2	فَعَلَ	فَعَلَ	2	يَنْصُرُ	نَصَرَ	2
ف ع ل	فَعَلَ	3	فَعَلَ	فَعَلَ	3	يَذْهَبُ	ذَهَبَ	3
ف ع ل	فَعِلَ	4	فَعِلَ	فَعِلَ	4	يَمْرَضُ	مَرِضَ	4
ف ع ل	فَعِلَ	5	فَعِلَ	فَعِلَ	5	يَحْسِبُ	حَسِبَ	5
ف ع ل	فَعُلَ	6	فَعُلَ	فَعُلَ	6	يَكْرُمُ	كَرُمَ	6
. .								

لائحة الأوزان

1	فَعَلَ	فَعَلَ	فعلت	فَعَلْتُ	فَعَلْتُ	1	
1	فَعَلَ	فَعَلَ	فعلنا	فَعَلْنَا	فَعَلْنَا	2	2
1	فَعَلَ	فَعَلَ	فعلنا	فَعَلْنَا	فَعَلْنَا	3	3
1	فَعَلَ	فَعَلَ	فعلت	فَعَلْتَ	فَعَلْتَ	4	4
. .							

قاموس الأوزان المصرفة

ج) لائحة السمات الصرف-تركيبية الجوهرية للأفعال العربية التي تشمل سمة الزمن والبناء والوجه والإعراب والسمات التطابقية (العدد، والجنس، والشخص).

د) لائحة تضمن جميع اللواصق (السوابق واللواحق) الممكن تأليفها مع الأفعال في اللغة العربية، والتي نميز بين نوعين منها: أولهما نسميه باللواصق غير المباشرة التي لا تؤثر مباشرة في تأويل

3.1. مقاربة التحليل الصرفي

تقوم هذه المقاربة بالبحث عن الجذع أو الجذر المناسب للكلمة السطحية التي يتم التعامل معها بعد إزالة كل اللواصق العالقة بها، مستعملة لهذه الغاية تقنيات مختلفة وأنظمة متباينة. ولعل ما يميزها عن سابقتيها كونها حاولت فهم عدد كبير من الظواهر المعتادة في الصرف العربي مستغلة بعض النتائج القيمة التي تم التوصل إليها في هذا المضمار، مما أضفى عليها نوعا من المعقولية في التعامل مع الكلمة العربية ولواصقها المختلفة. فبرزت للوجود مجموعة من المحللات الصرفية التي نخص بالذكر منها المحلل الصرفي لدرويش المعروف باسم "سيبويه" (Darwish, K, 2002) الذي يعتمد الوزن ليبحث عن الجذر والمحلل الصرفي لخوجة، والمحلل الصرفي لبكوالتر (Buckwalter, T., 2004) الذي يعتمد قاموسا يدويا ويستهدف الجذع والمحلل الصرفي لدياب، والمحلل الصرفي للسعداني وحشيش وغيرها.

2. نظام التحليل الصرفي المقترح

يشتغل هذا المحلل بكيفية تدريجية منطلقا من النص العربي الذي يُجزئ، طبقا للفراغات التي تترك بين الكلمات أثناء الكتابة، بواسطة مجزئ مهيأ لهذه الغاية، لنحصل في آخر هذه العملية على لائحة الكلمات التي يتضمنها هذا النص. بعدها يقوم النظام بتفكيك الكلمة السطحية أو النصية إلى عناصرها الذرية، وللتمييز بين أقسامها من فعل واسم وحرف يعتمد النظام على لائحتي السوابق واللواحق التي تلحق بكل منهما على حدى مع التأكد من أن الأحرف الأولى من الكلمة بالفعل هي أحرف زائدة وليست أصلية. فإن كانت الكلمة اسما فسيتم تحليلها عن طريق قالب تحليل الأسماء أما إذا كانت فعلا فسيتم تحليلها بواسطة قالب تحليل الأفعال.

1.2 قالب معالجة الأفعال

يقوم النظام على تحليل الكلمات السطحية وتفكيكها إلى سابقة وجذع ولاحقة، معتمدا في ذلك على موارد لسانية محصورة، ومصرف آلي (تم تطويره محليا) وتتمثل أساسيات هذا القالب:

1.1.2 الموارد اللغوية لقالب تحليل الأفعال

يستمد النظام المتبنى فعاليته من قاعدة معطيات رباعية، السواد الأعظم منها آلي والجزء اليسير منها يدوي. وقد تم تبني مقاربة قائمة على استخلاص القواعد من قاعدة المعطيات المشكلة، وهي تتضمن ما يلي:

أ) قاموس للأفعال العربية يتضمن حوالي 1000 فعل عربي مرقمة بحسب وزنها في قاموس الأوزان الذي يشكل نواة هذه القاعدة.

أ) حذف الحروف الثلاثة الأولى إذا كانت موجودة في لائحة السوابق الخاصة بها عندما يكون طول الكلمة خمسة أحرف على الأقل،

ب) حذف الحرفين الأولين إذا كانا موجودين في لائحة السوابق الخاصة بهما عندما يكون طول الكلمة أربعة أحرف على الأقل،

ج) حذف الحرف الأول "و" إذا كان طول الكلمة أربعة أحرف على الأقل وتبدأ ب "و"،

د) حذف الحرف الأول "ب" أو "ل" إذا كان طول الكلمة أربعة أحرف على الأقل وتبدأ ب "ب" أو "ل"، وإذا كانت موجودة في مجموعة الوثائق العربية المعتمدة،

ه) الإزالة المتكررة للاحقتين المواليتين في اللائحة المناسبة إذا كان طول الكلمة أربعة أحرف على الأقل قبل حذف هذه اللاحقة،

و) الإزالة المتكررة للاحقة الأولى الموالية في اللائحة المناسبة إذا كان طول الكلمة ثلاثة أحرف على الأقل قبل حذف هذه اللاحقة.

2.1. مقاربة حذف اللاصقة

تهدف مقاربة حذف اللاصقة، وهي المعروفة أيضا في الأدبيات باسم التجذيع الخفيف (light stemming)، إلى حذف السوابق واللواحق (دون الأواسط) من الكلمات السطحية/النصية العربية بغية استخلاص جذوعها من دون أن تهتم بجذورها الممكنة أو أوزانها المحتملة. من بين الباحثين الذين اشتغلوا في إطار هذه المقاربة وحاولوا تطبيقها على اللغة العربية، نذكر على سبيل المثال لا الحصر كلا من لاركي وآخرين.

التجذيع الخفيف حسب لاركي ((Larkey, L. S. et al, 2002 لا يتقيد بوجود قاموس للكلمات العربية، كما أنه لا يتقيد بالمقياس القائل بإمكان حذف اللاصقة عندما تكون البقية كلمة موجودة في اللغة العربية. ويقوم مجذع لاركي على ما يعرف بخفيف10 (light10) الذي لا يتوفر على جميع اللواصق الموجودة في اللغة العربية، بل يتضمن المستعمل منها بكثرة بحسب قياس ترددها في المتن المعتمد، ويعتمد على الخطوات الآتية:

أ) حذف "و" من خفيف2 وخفيف3 وخفيف8 إذا كان ما تبقى من الكلمة هو ثلاثة أحرف أو أكثر،1

ب) حذف أداة التعريف إذا كان ما تبقى من الكلمة هو حرفان،

ج) العودة إلى لائحة اللواحق وحذف أية لاحقة موجودة في آخر الكلمة إذا كانت هذه العملية تبقي حرفين أو أكثر.

التي كانت تروم بلوغ الجذر بدل الاكتفاء بالجذع، كما أن ظاهرة جمع التكسير (غير القياسي) شكلت عقبة صعبة حالت دون التوصل إلى نتائج وحلول دقيقة، هذا علاوة على قضية تخلي الكتابة العربية عن الحركات القصيرة وما يخلفه من لبس على مستوى الكلمة في حد ذاتها، مما يعيق عمل بعض الأنظمة التي تأخذ بعين الاعتبار جوانب معينة.

وإذا ألقينا نظرة شاملة على جل هذه الأعمال يمكن تلخيص مضامينها باعتماد المقابلات أو الثنائيات التالية:

- أعمال تستهدف الجذع وأخرى تستهدف الجذر،

- أعمال تعتمد على قواميس وأخرى لا تعتمد عليها،

- أعمال تستعين بالأوزان وأخرى لا تستعين بها،

- أعمال تلجأ إلى الترجمة الآلية (المقابلات الإنجليزية للكلمات العربية) وأخرى لا تلجأ إليها،

- أعمال قائمة على لائحة تتضمن عددا أكبر من عدد اللواصق الموجودة في اللغة العربية وأخرى قائمة على لائحة تتضمن البعض منها فقط،

- أعمال تدرج معلومات صرف-تركيبية عن عناصر أو مكونات الكلمة النووية وأخرى لا تقدم عنها أية معلومة على الإطلاق.

جل هذه الأعمال تندرج بكيفية مباشرة أو غير مباشرة في إحدى المقاربات الثلاث الموالية التي تدور في فلكها معظم الأبحاث الحديثة التي تسهم بشكل فعال في تطوير هذا الميدان: المقاربة الإحصائية، ومقاربة حذف اللاصقة، ومقاربة التحليل الصرفي، وهي التي سنركز عليها في هذه الورقة.

1.1. المقاربة الإحصائية

جوهر مقاربة التجذيع الإحصائي هو حساب ترددات توارد الجذوع واللواصق، وتعتمد على نماذج n-grams لاستخراج الجذوع من الكلمات السطحية. يتم تكوين طبقات من الكلمات التي تبدأ بالحرف الأول نفسه (n-grams)، أو الجزء الأول منه، واستعمال تقنيات التصفيف لتشذيب هذه الطبقات. إنها مقاربة تعتمد على التحليل الصرفي الآلي وعلى تحليل التوارد الذي يقيم طبقات من الجذوع على أساس قياس التوارد. وقد طبقت هذه المقاربة على اللغة العربية من قبل (Xu J., et al, 1998) و (Mayfield J., et al., 2003) و (De Roeck, A. N., et al, 2000).

ينبني مجذع شين (Chen, A.,) ، مثلا، على الترجمة الآلية (إنجليزي/عربي)، إذ يستخلص الجذع من الكلمات العربية بناءا على مقابلاتها الإنجليزية. فالمقابل الإنجليزي لكلمة عربية من قبيل أطفال هي children، وبما أن المجذع الإنجليزي له القدرة على تحويل الأسماء المجموعة إلى مفرد عن طريق إزالة لاصقة الجمع (نظرا للطبيعة الصرفية لهذه اللغة)، فإنه سيعثر على كلمة child التي تقابلها في اللغة العربية كلمة طفل. بهذه الطريقة يتم التوصل إلى جذوع الكلمات العربية بما في ذلك مفرد جمع التكسير. لقد قام شين بتطوير مجذع خفيف يقوم بحذف وإزالة السوابق واللواحق المترجمة إلى اللغة الإنجليزية، واضعا ست لوائح تتضمن الثلاث الأولى منها الحرف الأول والحرفين الأولين والحروف الثلاثة الأولى، بينما تشتمل اللوائح الثلاث الأخرى على الحرف الأخير والحرفين الأخيرين والحروف الثلاثة الأخيرة. الخطوات المتبعة في هذا المجذع هي:

محلل صرفي عربي للنصوص العربية

عبد الفتاح حمداني، سعيد الحسني

معهد الدراسات والأبحاث للتعريب، الرباط
said.elhassani@gmail.com
fattahhamdani@gmail.com

مقدمة

تعتبر المحللات الصرفية إحدى اللبنات الأساسية لبناء جل التطبيقات في ميدان المعالجة الآلية للغات الطبيعية وقد تم اقتراح عدد لا بأس به منها في غضون السنوات القليلة الأخيرة من قبل فرق ومختبرات بحث وباحثين عرب وغير عرب ونذكر منهم كلا من باكوالتر (Buckwalter, T., 2004) ولاركي ((Larkey, L. S. et al, 2002 وخوجة (Khoja, S., Garside, 1999) ودرويش ((Darwish, K, 2002 ودياب والسعداني وحشيش وغيرهم كثير .

نبدأ هذه الورقة بعرض لبعض النماذج من المقاربات المعروفة لإنجاز المجذعات (stemmers) والمحللات الصرفية، حيث سنبدي بعض الملاحظات العامة والموجزة على المنهجية المتبعة فيها، ثم نقدم المنهجية التي نروم اعتمادها لإنجاز محلل صرفي عربي يراعي عددا غير قليل من الخصوصيات التي تنفرد بها اللغة العربية عن غيرها من اللغات الطبيعية الأخرى، وبناءا على أقسام الكلمة (اسم، فعل، حرف) تم تقسيم النظام إلى قالبين رئيسيين حيث سنقدم تصورنا لقالب تحليل الأسماء، بدء بتصنيفها، وانتهاء بالسمات التي ترتبط بها، وبعدها سنتطرق لقالب تحليل الأفعال.

1. مقاربات جوهرية ومحللات صرفية للغة العربية

يعد مفهوم التجذيع (stemming) في مجال اللسانيات الحاسوبية مفهوما جوهريا قامت على أساسه العديد من المجذعات والمحللات الصرفية التي تكثر الحاجة إليها في جل التطبيقات اللسانية الحاسوبية. والمقصود بالتجذيع، بصفة عامة، عملية حذف أو إزالة اللواصق من الكلمات السطحية/النصية للحصول على الجذع أو الجذر. وقد نالت الكلمات العربية قسطا لا بأس به من الاهتمام عكسته الأعمال الكثيرة التي أنجزت في هذا الإطار، والنتائج القيمة التي تمكنت من تحقيقها في فترة وجيزة للغاية.

ويكاد يجمع جل الباحثين في هذا الحقل المعرفي الفتي على أن صرف اللغة العربية يتطلب إجراءات إضافية لكونه صُرفيا واشتقاقيا في الآن نفسه. فمشكل الأواسط طرح بحدة خصوصا في الأعمال

PROSEM et gestion de la sémantique contextualisée
Quelques domaines d'application

Hammou Fadili[1, 2]

[1]Pôle recherche de la Maison des Sciences de l'Homme de Paris
190 avenue de France 75648 Paris Cedex 13, France
hammou.fadili(at)msh-paris.fr

[2]Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
hammou.fadili(at)cnam.fr

Résumé

Le contexte est mal géré dans les domaines où il constitue un élément très important et parfois même incontournable, comme dans le domaine du Traitement Automatique des Langues (TAL), dans l'analyse sémantique de données et dans la gestion de connaissances (Semantic Data Mining & Knowledge Management : SDMKM), etc. Pour remédier à ce problème, nous avons proposé une approche appelée « PROSEM » (PROjection SEMantique) permettant de détecter et de relever tous les « traits sémantiques » par rapport à un contexte donné et d'augmenter largement l'efficacité de la formalisation du sens d'un contenu afin qu'il soit compris par la « machine ». De par sa nature, cette projection peut être utilisée dans beaucoup d'autres domaines liés à la notion du contexte comme : l'indexation des données, la recherche linguistique ou sémantique dans les systèmes d'information, l'extraction contextuelle d'information, la gestion des corpus multilingues, la génération automatique de textes, etc. Dans cet article, nous allons présenter quelques domaines d'application qui peuvent être améliorés par PROSEM ainsi qu'une évaluation des mesures liées aux performances du modèle proposé.

Mots clés : *PROSEM, TAL, SDM, KM, Contexte, Sémantique, Ontologies, Indexation, Recherche d'Information (RI), Extraction d'Information (EI).*

1. Introduction

Le contexte est mal géré dans les domaines où il constitue un élément fondamental, comme dans le domaine du Traitement Automatique des Langues (TAL), dans l'analyse sémantique de données et dans la gestion de connaissances (Semantic Data Mining & Knowledge Management : SDMKM), etc. Pour remédier à ce problème, nous avons proposé une approche appelée PROjection SEMantique « PROSEM » (H.FADILI et M.CHAKIRI (SITACAM 2009)) permettant de détecter et de relever tous les « traits sémantiques » par rapport à un contexte donné et d'augmenter l'efficacité de la formalisation du sens d'un texte afin qu'il soit compris par la « machine ».

La modélisation de cette « relation, association » entre le texte et le contexte basée sur PROSEM a pour but de formaliser le sens par l'extraction des mots et des relations permettant l'obtention d'un réseau sémantique « *contextualisé* » reflétant fidèlement le sens des contenus étudiés. Concrètement, elle permet de faire « *la projection* » du texte représenté par l'arbre conceptuel (AC) issu du TAL sur le contexte représenté par l'ontologie du contexte (OC) en utilisant les « *relations sémantiques* » pour faire le « *mapping* » entre les concepts, mots, relations, instances, attributs, etc. des graphes. Nous avons tenu à mettre l'accent sur l'utilisation des relations sémantiques parce que nous considérons que c'est le moyen le plus sûr permettant d'éviter la déperdition du sens et de relever toutes les nuances dans un contenu en rapport avec le contexte. Nous essayons à travers cette analyse de prendre en compte certaines relations sémantiques (un sous-ensemble de l'ensemble de toutes les relations), celles que nous avons souhaité utiliser dans un premier temps afin de valider notre approche. Nous faisons référence ici aux relations de synonymie, de polysémie, d'homonymie, d'antonymie, d'hyperonymie, d'hyponymie, etc. En effet, nous considérons que si un mot est important et qu'il doit être retenu dans un contexte, il en est de même pour ces synonymes, ses contraires, ses génériques, ses spécifiques, etc.

La spécificité de chaque type de relations sémantiques est gérée dans les différents algorithmes de la projection. Bien évidemment, l'extension vers d'autres relations est possible suivant le même principe. Outre les mots simples et les phrases ordinaires, nous avons étendu notre approche à des tournures figées et métaphoriques. Elles sont codées ou indexées pour que la machine puisse les extraire facilement et leur réserver un traitement spécial pour éviter toute ambigüité au niveau de l'interprétation. Concrètement, on les exclut des traitements du TAL pour les remplacer par les représentations formelles de leurs sens réels dans l'arbre

final. De par sa nature, PROSEM peut être utilisée dans beaucoup de domaines liés à la notion du contexte comme : l'indexation des données, la recherche linguistique ou sémantique d'information, l'extraction contextuelle d'information, la gestion des corpus multilingues, la génération automatique de textes, etc. Dans cet article, nous allons présenter quelques domaines d'applications de PROSEM ainsi que l'apport que cette méthodologie pourrait apporter pour augmenter l'efficacité dans ces domaines.

2. Quelques éléments sur PROSEM

Ce paragraphe sera consacré à un rappel sur quelques éléments essentiels de PROSEM. Pour plus d'informations, cf. « PROSEM » (PROjection SEMantique) H.FADILI et M.CHAKIRI (SITACAM 2009). Le processus consiste, à intégrer des outils permettant d'extraire l'arbre sémantique brut en utilisant les outils du TAL, à le convertir en un arbre conceptuel en appliquant les algorithmes de reconnaissance et de classification d'entités et de résolution d'anaphores pour effectuer la « *projection sémantique* » de l'arbre obtenu sur l'ontologie du contexte « *sémantiquement parallèle* » aux relations sémantiques décrites dans le paragraphe suivant.

2.1. Les relations sémantiques

Les relations sémantiques représentent les liens de sens que peuvent entretenir deux ou plusieurs mots par rapport à leurs significations, comme les relations de type synonymie, antonymie, hyperonymie, etc. Dans les cas des applications traitées dans cet article, nous pensons que l'utilisation de ces relations est importante et permet, par exemple, d'éviter d'ignorer des mots bien qu'ils soient pertinents par rapport au contexte. On considère que si un mot est pertinent dans un contexte, alors tous les mots que l'on peut atteindre via les relations sémantiques sont aussi importants que le mot lui même.

Etant donné que le nombre de relations sémantiques est important, la probabilité qu'un mot et les mots sémantiquement liés soient considérés est supérieure à la probabilité ne tenant compte que du mot seul. L'utilisation de ces relations peut avoir un impact très positif sur l'amélioration des performances des applications où elles peuvent être utilisées.

Dans cet article, nous nous sommes intéressés à un sous-ensemble de l'ensemble des relations sémantiques entre les mots pour montrer l'apport que pourrait apporter la démarche PROSEM à travers cet échantillon. A ce stade de la

recherche, on s'est limité à un traitement et à une utilisation générique des relations sémantiques, mais une étude très approfondie des spécificités de chaque relation ainsi que l'utilisation qu'on pourrait en faire est prévue dans nos travaux futurs.

Ci-après une présentation très brève d'une liste non exhaustive de relations sémantiques choisies qu'on pourrait traiter de la même manière dans le cadre de cette démarche :

Etymologie : origine ou filiation d'un mot.

Synonymes : mot, syntagme qui par son sens est similaire à un autre. Exemple : grand / élevé/.

Antonymes : mot, syntagme qui par son sens s'oppose à un autre. Exemple : grand / petit.

Hyperonyme : mot (nom) dont le sens inclut d'autres mots (générique). Exemple : « insecte » est l'hyperonyme de « abeille ».

Hyponyme : terme désignant une sous-classe. Exemple : « abeille » est l'hyponyme d'« insecte ».

Méronymes : Terme lié à un autre par une relation de partie à tout. Exemple : Voiture / roue.

Troponymes : Pour les verbes, qui décrivent de manière plus précise la façon dont l'action d'un autre verbe (le verbe dont il est le troponyme) est réalisée. Exemple : Se déplacer / marcher.

Dérivés : sont des mots obtenus à partir de la même racine.

Expressions figées : Une locution est toute suite polylexicale construite d'unités lexicales non soudées, formant un bloc figé inanalysable au niveau sémantique. (Chakiri 2007).

2.2 Fonctions fondamentales

Ci-après la définition de quelques fonctions que nous avons utilisées dans le contexte de PROSEM et qui ont aussi servi pour la définition des extensions des domaines d'applications traités dans cette étude.

<u>Soit la fonction « Relations Sémantiques » RS d'un mot</u> :

$$(Di) \longrightarrow \{\{m\}\}$$

$$Xi \longrightarrow RS(Xi)$$

Tel que : $RS(Xi) = \{Xi/syn(Xi)/ant(Xi), \ldots\}$

(Di) un document et {{m}} ensemble d'ensemble de mots.

Soit la fonction « Lien Sémantique » LS entre mots :

$$(Di)*(Dj) \longrightarrow \{0,1\}$$

$$(Xi,Xj) \longrightarrow LS(Xi,Xj)$$

Tel que : $LS(Xi,Xj) = 1 \text{ si } RS(Xi) \cap RS(Xj) \neq \phi$

$LS(Xi,Xj) = 0 \text{ sinon}$

(Di)et (Dj)sont des documents.

Soit la fonction « Intersection Sémantique » IS entre documents :

$$\{Di\}*\{Dj\} \longrightarrow \{0,1\}$$

$$(di,di) \longrightarrow IS(di,di)$$

Tel que : $IS(di,di) = 1 \text{ si } \exists Xi \in di, \exists Xj \in dj \,/\, LS(Xi,Xj)=1$

$IS(di,di) = 0 \text{ si } \forall Xi \in di, \forall Xj \in dj \,/\, LS(Xi,Xj)=0.$

{Di} et {Dj} sont des ensembles de documents.

Soit la fonction « Projection Sémantique » PS par rapport à une ontologie du domaine (OD) :

$$(Di) \longrightarrow \{m\}$$

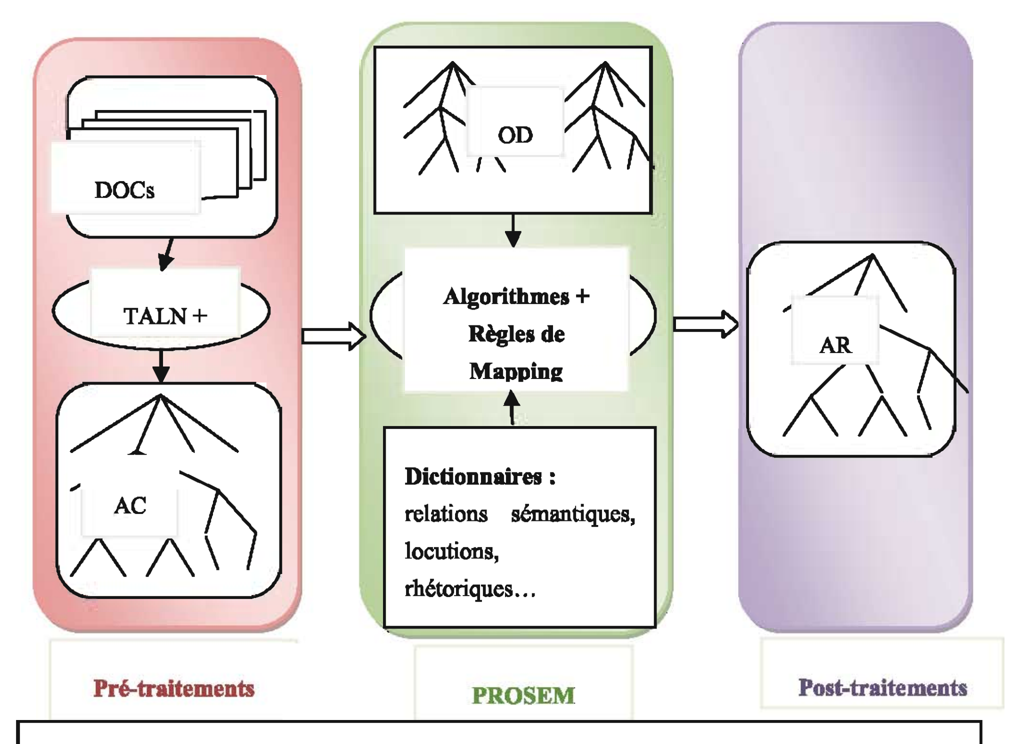
$$Xi \longrightarrow PS(Xi)$$

Tel que : $PS(Xi) = m \in RS(Xi) / \exists n \in RS(Xi) \text{ et } n \in (OD)$, *m est choisi en fonction de l'utilisation (index, recherche, extraction, ...)*

$PS(Xi) = O \text{ sinon}$

(Di) un document.

2.3 Architecture



Légende : DOCs : Documents, TAL : Traitements Automatiques des Langues Naturelles, EN : entités nommées, AC : Arbre Conceptuel, OD : Ontologie du Domaine, AR :

Figure 1. Architecture générale de PROSEM

Pour expliciter et expliquer le schéma précédent, nous présentons l'algorithme principal décrivant le processus PROSEM.

2.4. Algorithme principal

Algorithme principal

Entrée *: ensemble de documents (D), ontologie du domaine et des contraintes (ODC), arbre conceptuel (AC), dictionnaire d'entités nommées (EN), dictionnaire des relations conceptuelles(DIC), dictionnaire des locutions (LOC), objets du discours (DO).*

Sortie *: arbre conceptuel « compris »*

Début

(AG) = ∅ -- arbre généré final vide

***Pour** chaque document **(Di)** faire*

(AGi) = ∅ -- arbre généré pour chaque document

*chaque nœud (Ni) ∈(AC) **faire***

***Action 0. si** (Ni) ∈(ODC) **alors** (AGi) ++ (Ni)*

***Action 1. si** syn(Ni) ∈(ODC) **alors** (AGi) ++ (Ni)*

Action 2. si** ant(Ni) ∈(ODC) **alors

***Action 3.** chercher les relations associées*

***Action 4.** (Agi) ++ action inverse (ant(Ni)) -- A a vendu à B /B a acheté à A*

***Action 5. si** hypo(Ni) ∈(ODC) **alors** (AGi)++(Ni)*

Action 6. si** hyper(Ni) ∈(ODC) **alors

***Action 7.** Lire et appliquer le choix de l'utilisateur*

***Action 8.** Répéter pour toutes les relations sémantiques*

Chaque** locution (LOCi) E (Di) **faire

***Action 9.** (Agi)++ graphe (LOCi)*

***Fin** chaque*

Chaque** locution (DO) E (Di) **faire

***Action 9.** (Agi)++ graphe (DO)*

***Fin** chaque*

(AG)++(Agi)

***Fin** pour*

***Retour** (AG)*

Fin

Les différents domaines d'application que l'on présente dans cet article et que l'on veut améliorer et enrichir avec PROSEM peuvent être classifiés suivant trois axes,

un axe orienté vers l'indexation, un axe qui traite de la Recherche d'Information (RI) et puis enfin un axe orienté vers l'Extraction d'Information (EI). Dans ce qui suit, nous allons définir brièvement chaque axe, puis montrer comment l'on peut à travers les fonctionnalités de PROSEM améliorer l'optimisation ainsi que la pertinence des résultats d'utilisation.

3. Processus d'indexation contextuel

Un index est une vue des éléments (contenus) qu'il représente, base de données, documents etc. Nous pouvons parler par exemple d'un index des auteurs, des villes ou d'un index d'une catégorie de mots particuliers pour une utilisation donnée. L'indexation consiste en la construction des tables décrivant *certaines* données ainsi que leurs emplacements dans les systèmes où elles sont stockées, ceci permet de faciliter leur localisation et leur mode d'accès, surtout nécessaires dans le cas des systèmes d'information volumineux.

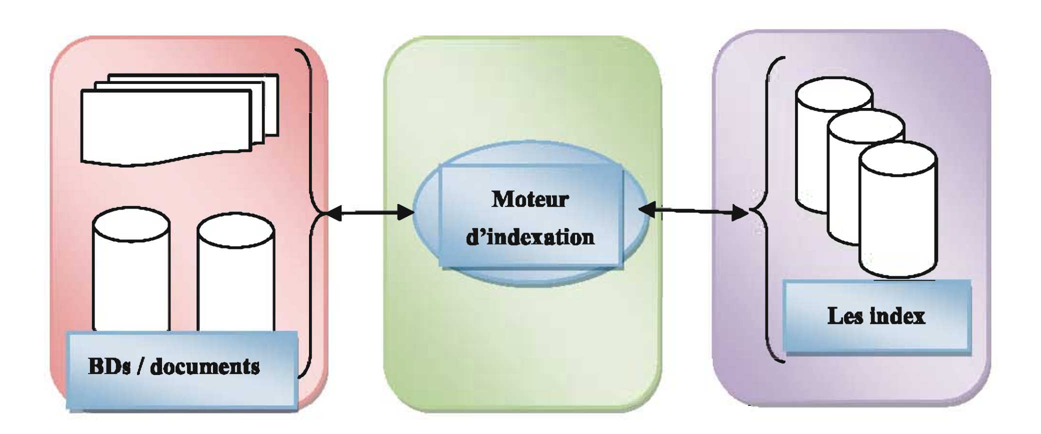


Figure 2. Processus général d'indexation

La plupart des moteurs d'indexation actuels génèrent souvent des index redondants et parfois non pertinents. Ceci est dû au fait que les entrées sont souvent générées par des algorithmes simples considérant chaque mot du corpus comme entrée potentielle dans l'index. On fait correspondre à chaque mot une entrée dans l'index, considérant par exemple les mots ayant le même sens comme différents, c'est l'égalité stricte entre les mots qui est utilisée pour alimenter l'index suivant le processus ci-après :

Si le mot possède déjà une entrée dans la table des index, on procède à une mise à jour (une sorte d'UPDATE) de l'enregistrement en y rajoutant l'emplacement du mot en question, sinon on crée une nouvelle entrée dans la table des index avec le mot en question et son emplacement.

Algorithme du processus classique d'indexation

Entrée:
- ensemble de documents ou éléments à indexer (ED)
Sortie:
- Index généré (IG).

Début

Pour** chaque (ED) **faire :

Pour** chaque élément (Xi) de (ED) **faire :

***- Action 0. Si** Xi € (IG)*

***- Action 2. Alors** UPDATE ((IG), Xi)*

- Action 3. Sinon

***- Action 4.** INSRT((IG),Xi)*

- Action 6. Fin si

Fin pour

Ces types de processus ont plusieurs inconvénients, malgré qu'un index est une vue condensée du contenu qu'il représente, l'indexation des systèmes de gros volumes de données peut générer des quantités très importantes d'information, parfois difficiles à gérer et à maintenir. En effet, la taille d'un index et des traitements associés augmentent proportionnellement en fonction de la taille des données. Un autre inconvénient qui caractérise les index issus des techniques dites classiques, est leur pertinence. La plupart des index sont générés de la même manière avec les mêmes algorithmes, sans tenir compte du contexte et du but de l'utilisation. L'indexation universelle ou celle qui répond à toutes les utilisations n'existe pas, du fait, qu'il existe un nombre infini de vues que l'on peut associer à un contenu. Une indexation doit être faite dans le but de répondre à une utilisation donnée par la délimitation du domaine et la définition du contexte d'utilisation ; ce qui permet de prendre en charge les problématiques d'indexation liées essentiellement à la pertinence et l'optimisation (non redondance au sens sémantique).

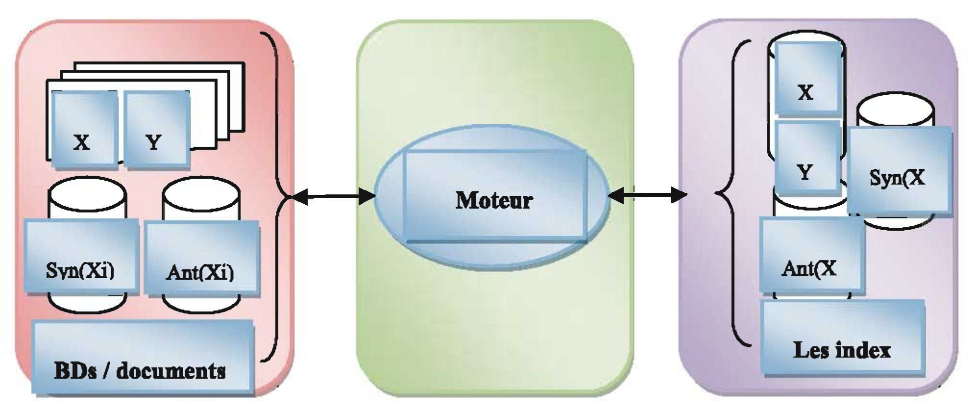


Figure 3. Processus classique de gestion/génération d'index,

Ces problèmes sont dus en partie au fait que l'analyse du contenu est basée essentiellement sur des méthodes simples traitant chaque élément, indépendamment des autres et en dehors du contexte, même si des « factorisations » sont possibles pour simplifier et diminuer la complexité et la quantité d'information à retenir. C'est pour cela que l'on propose d'utiliser l'approche PROSEM afin d'enrichir l'analyse du corpus dans le contexte pour générer un index optimal.

Algorithme du processus général d'indexation en utilisant PROSEM

Entrée *:*
- ensemble de documents ou éléments à indexer (ED)
- ontologie du domaine et des contraintes (ODC).

Sortie *:*
- Index généré (IG).

Fonctions spécifiques *:*
- UPDATE ((IG), Xi) : permet de mettre à jours l'entrée Xi ou RS(Xi) dans l'index en avec l'emplacement de Xi.
- INSERT((GI),Xi) : permet de créer une nouvelle entrée Xi dans l'index avec son emplacement.
- Les fonctions (RS) et (PS) définies précédemment.

Début

Pour *chaque (ED)* ***faire*** *:*

Pour *chaque élément (Xi) de (ED)* ***faire*** *:*

- Action 0. Si ∃ m € RS(Xi) / m € (ODC)

- Action 1. Si ∃ m € RS(Xi) / m € (IG)

- Action 2. Alors UPDATE ((IG), Xi)

- Action 3. Sinon

- Action 4. INSRT((IG),Xi)

- Action 5. Fin si

- Action 6. Fin si

Fin pour

Cette nouvelle approche d'indexation permet d'introduire l'analyse sémantique du « contenu » en introduisant la notion du domaine et du contexte pour générer l'index. Cette approche permettra d'obtenir un index par domaine et/ou par centre d'intérêt, pertinent et optimisé. Ci-après la description du processus général de la nouvelle approche.

Nous vérifions pour chaque mot l'existence dans l'ontologie du domaine et des contraintes ; l'existence du mot dans l'ontologie est une existence au sens PROSEM qui la définit comme suivant : un mot est dans l'ontologie (ou fait partie du domaine traité), si le mot existe dans l'ontologie ou si au moins une de ses images *« RS(mot) »* par les relations sémantiques existe dans l'ontologie. Plus concrètement, si le mot existe dans l'ontologie au sens PROSEM, on fait un UPDATE de l'index de l'entrée avec l'emplacement de l'image *« RS(mot) »* présente dans l'ontologie, sinon on crée une nouvelle entrée dans la table des index en utilisant un élément de *« RS(mot) »* présent dans l'ontologie (le nœud représentant les éléments de *RS(m))* comme entrée de l'index.

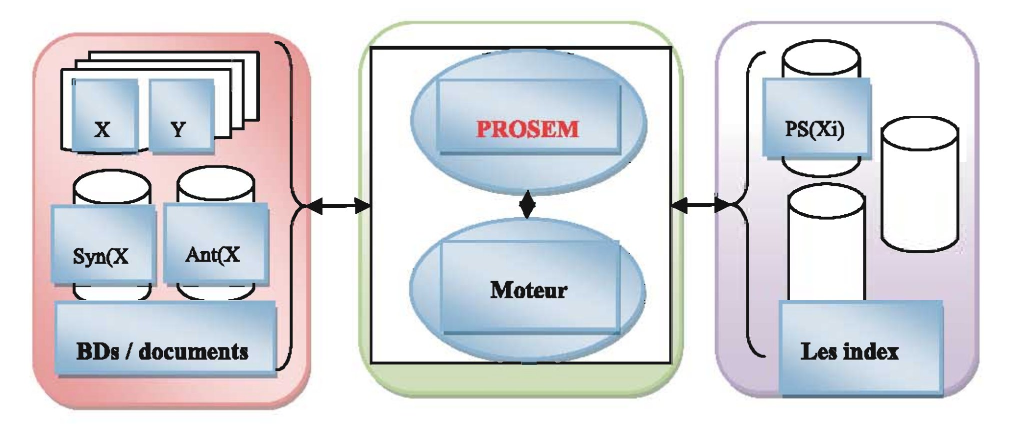


Figure 4. Processus de gestion/génération d'index en utilisant PROSEM

Avec cette méthode, l'indexation est optimisée par rapport à ce qu'elle aurait été avec les méthodes classiques. Elle est pertinente et répond à une utilisation dans un contexte donné et la quantité d'informations stockée est largement inférieure. Elle peut être estimée à *« (taille de l'index / (le nombre de relations sémantiques)) – le Bruit »*. En effet, et d'une part, tous les mots liés par les relations sémantiques sont représentés par une seule entrée dans l'index, et d'autre part, tous les mots ne faisant pas partie du domaine sont éliminés. Les mesures de pertinence et de quantité d'information qu'on obtient avec PROSEM sont prévues dans une étude ultérieure.

4. Recherche sémantique et contextuelle d'information

Les moteurs de recherche simples permettent de retrouver des occurrences de mots dans des documents et dans des bases de données ainsi que leurs emplacements puis renvoyer à l'utilisateur les contenus contenant les occurrences recherchées.

Les recherches classiques consistent à comparer les éléments de la requête avec ceux des textes sans tenir compte d'aucune contrainte, comme le contexte, le domaine ou autres. Ce sont des comparaisons booléennes ne permettant pas souvent d'obtenir des résultats satisfaisants.

Ils existent aussi ce que l'on appelle des moteurs dits « avancés », la plupart fonctionnent suivant le processus suivant : la requête demandée par un utilisateur est analysée puis décomposée en plusieurs termes et opérateurs booléens. Les recherches consistent à trouver, puis à renvoyer à l'utilisateur les emplacements des

contenus contenant les termes de la requête par combinaison des opérateurs booléens de la requête.

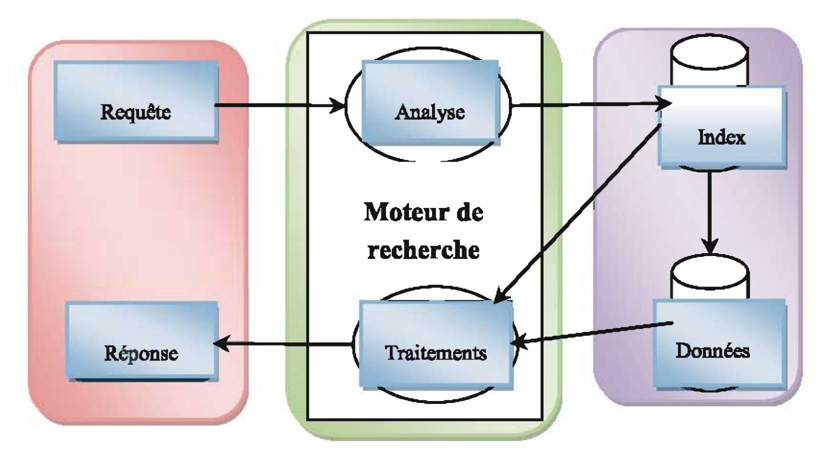


Figure 5. Processus classique de recherche

Malgré leurs performances, les résultats fournis par ce type de moteurs sont dans la plupart des cas incomplets et impertinents, c'est-à-dire, que l'on peut avoir des résultats avec beaucoup de « bruit » comme on peut y ignorer beaucoup d'informations pertinentes « silence ». C'est pour cela que nous proposons de coupler le mode de fonctionnement de ce type de moteurs avec les techniques issues de PROSEM pour apporter des améliorations : au niveau de l'analyse des requêtes, au niveau de l'analyse des contenus et aussi au niveau de la recherche, c-à-d. au niveau du « mapping » entre les requêtes et les contenus. Ceci suivant la démarche ci-après :

Au lieu de chercher la liste des mots clés de la requête directement dans les index, nous procédons tout d'abord à une analyse sémantique de la requête pour générer l'équivalent d'un arbre conceptuel (AC) décrit précédemment, puis nous faisons la projection de l'arbre obtenu sur l'ontologie du domaine et des contraintes (ODC) et enfin nous faisons des comparaisons avec l'index en utilisant les relations sémantiques pour déduire les résultats de la requête. Cette démarche a l'avantage de permettre de limiter le domaine de recherche, bien cibler les intensions des demandes de l'utilisateur et les messages véhiculés dans les contenus, afin d'augmenter les possibilités de recherche en se basant sur les relations sémantiques (RS). Ce qui permet d'éviter la déperdition du sens au niveau des demandes utilisateurs, au niveau des contenus analysés et aussi au niveau des liens entre eux.

Figure 6. Processus de recherche en utilisant PROSEM

Algorithme du processus général de recherche en utilisant PROSEM

Entrée:
- requête utilisateur (R)
- ensemble de documents à analyser (ED)
- ontologie du domaine et des contraintes (ODC).

Sortie :
- ensemble de contenus réponse(CR).

Fonctions spécifiques :
- Les fonctions (RS) et (LS),(IS), (PS) définies précédemment.

Début

Pour** chaque requête (R) **faire :
- Action 0. (CR) = ϕ

***- Action 1**. Construire l'arbre conceptuel (ACR) de (R)*
Pour** chaque contenu (ED) **faire :

***- Action 2**.charger l'arbre conceptuel (ACED)*

***- Action 3**. Faire la projection sémantique de graphes ACR'=PSG(ACR,ODC)*

***- Action 4**. Faire la projection sémantique de graphes ACED'=PSG(ACED,ODC)*

***- Action 5. Si** IS(ACR',ACED') = 1*

***- Action 6. Alors** (CR)++(ED)*

- Action 7. Fin si

***- Action 8. Retour** (CR)*

Fin pour

----

5. Extraction contextuelle d'informations

L'extraction d'Information consiste en l'identification des informations pertinentes dans un contenu pour une utilisation donnée. D'une manière générale, l'extraction d'information part d'un texte écrit en langue naturelle pour en extraire des informations souvent structurées sous forme de bases de données, d'index, d'ontologies, etc., pouvant respecter dans certains cas un schéma donné. Contrairement à la recherche d'information qui consiste à analyser les documents pour renvoyer à l'utilisateur les plus pertinents, l'Extraction d'Information (EI) analyse les documents pour ne renvoyer à l'utilisateur que les informations pertinentes.

La plupart des systèmes d'Extraction d'Information se basent grossièrement sur des tâches qui consistent à analyser et à extraire tous les mots d'un document. L'Extraction d'Information peut être utilisée dans plusieurs domaines comme :

- la reconnaissance d'Entités Nommées où les mots sont associés à des catégories.
- le peuplement d'ontologies où les mots sont associés à des concepts, attributs ou relations.
- Etc.

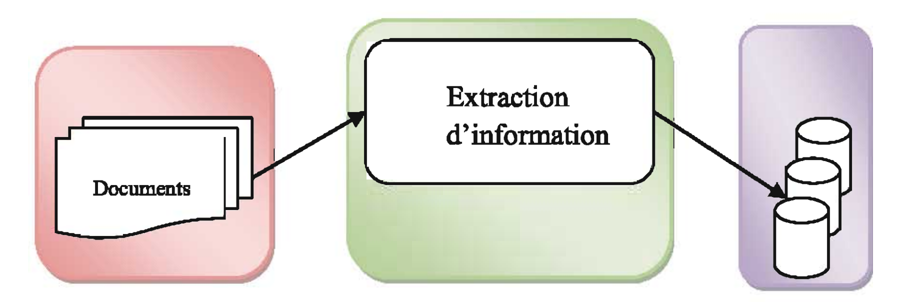


Figure 7. Processus classique d'extraction d'information

Les limites de ces démarches résident, comme dans les cas décrits précédemment, dans le fait que la comparaison des entités analysées se fait en utilisant la relation d'égalité booléenne indépendamment des domaines ou autres contraintes pour déterminer si une entité ou un terme doit être retenu « *extrait* » ou pas, ce qui est bien sûr insuffisant pour l'Extraction d'Informations pertinentes par rapport à un contexte.

Pour remédier à ce problème, nous proposons d'analyser la sémantique de la source d'information par rapport au contexte afin de mieux comprendre le sens exprimé et le message véhiculé pour ne retenir que les informations qui ont un sens dans le contexte en question. Nous pensons que la projection sémantique décrite précédemment (PROSEM) peut avoir un atout considérable pour mieux analyser et extraire les informations pertinentes. Plus concrètement, nous procédons à l'extraction de termes suivant une ontologie de domaine et des contraintes en utilisant les relations sémantiques décrites précédemment. Cette approche permet de faire un zoom sur cette relation de comparaison des unités d'informations pour retenir de nouveaux termes qui échapperaient aux techniques classiques et rejeter des termes qui seraient retenus.

Prenons le cas de la reconnaissance d'entités nommées par exemple, qui est effectivement une sous-tâche de l'extraction d'information. Elle permet de chercher et de classifier les éléments d'un texte (mots ou groupes de mots) suivant des catégories prédéfinies. Ces éléments peuvent être des personnes, des organisations, des dates, etc. Les systèmes actuels utilisent des dictionnaires « d'entités nommées » qui recensent tous les noms de personnes, toutes les villes, toutes les dates, organisations, etc. puis procèdent à la recherche/comparaison de tous les mots d'un document dans le dictionnaire. Si le mot ou groupe de mots existe dans le dictionnaire, nous lui associons la catégorie correspondante, sinon il n'est pas retenu dans la classification.

Exemple : Jean réside à Paris depuis 15/11/2000. En comparant avec un dictionnaire d'entité nommées, nous pouvons déduire que : Jean est une personne, Paris est une ville et le 15/11/2000 est une date.

Par contre, si nous rencontrons, dans un document, les mots : *un avocat véreux*. La recherche de ces mots dans un dictionnaire d'entité nommée dépend du domaine que l'on veut étudier. Les systèmes existants ne permettent pas de résoudre ce problème et cette ambigüité. C'est pour cela que l'on propose d'étendre les techniques existantes pour d'une part tenir compte du domaine pour la désambigüisation et d'autre part augmenter la probabilité de reconnaitre et classifier un maximum de mots d'un texte. Nous pensons que PROSEM peut jouer ce rôle de raffinement du traitement du sens des mots dans le processus de reconnaissances d'entités nommées.

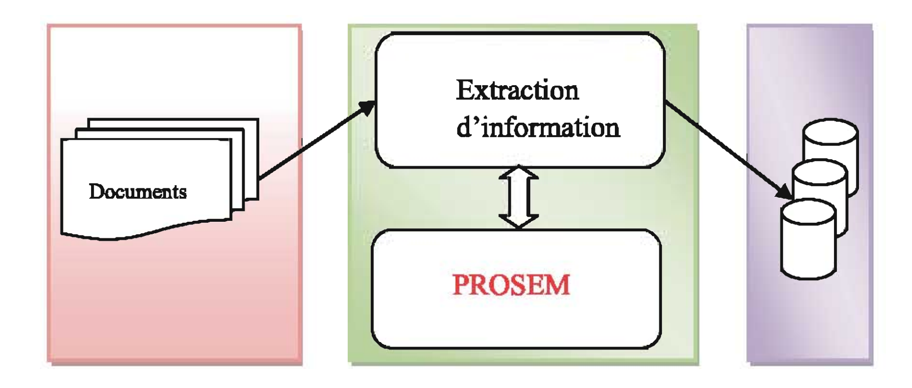


Figure 8. Processus d'extraction d'information en utilisant PROSEM

Les algorithmes du processus d'indexation décrits précédemment peuvent être améliorés et augmentés de modules d'hiérarchisation et de catégorisation pour la prise en charge de l'Extraction d'Information dans le contexte.

Algorithme du processus général d'Extraction d'Information en utilisant PROSEM

Entrée :
- ensemble de documents ou éléments à indexer (ED)
- ontologie du domaine et des contraintes (ODC).

Sortie :
- Information extraite (IE).

Fonctions spécifiques :
- INSERT((IE),Xi) : permet de créer une nouvelle entrée Xi dans l'ensemble des informations extraites.
- Les fonctions (RS) et (PS) définies précédemment.

Début

Pour** chaque (ED) **faire :

Pour** chaque élément (Xi) de (ED) **faire :

***- Action 0. Si** ∃ m € RS(Xi) / m € (ODC)*

***- Action 1.** INSERT((IE),Xi)*

- Action 2. Fin si

Fin pour

---

6. Quelques éléments pour l'évaluation des performances

Pour calculer et évaluer les performances de PROSEM, nous avons calculé et comparé les mesures de probabilité d'apparition d'un mot ou un groupe de mots dans un corpus ainsi que les mesures de probabilité que ce mot ou groupe de mots soient pertinents par rapport à une utilisation donnée, ceci en tenant compte ou pas de PROSEM.

La probabilité d'apparition d'un mot dans un document

Notre comparaison a été basée sur le calcul de probabilité d'apparition d'un mot dans un document. Cette probabilité peut être calculée de plusieurs façons et pour cela il existe des méthodes et techniques. Mais pour des raisons de complexité, nous avons choisi la plus simple, celle qui permet de calculer cette probabilité de la manière suivante.

Soient m , C et P_m, un mot m dans un corpus C et P_m la probabilité d'apparition de m dans C.

Alors : $P_m = \frac{Nb_{Occ(m)}}{Nb_{mots(C)}}$, où $Nb_{Occ(m)}$ est le nombre d'occurrences de m dans C,

et $Nb_{mots(C)}$ est le nombre de mots dans C.

La formule de Poincaré :

Cette formule consiste à calculer la probabilité d'union de plusieurs événements. Elle s'exprime :

Pour tout entier $n \geq 2$ et tous événements $\mathrm{A}_1,...,\mathrm{A}_n$:

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) + \sum_{k=2}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < i_2 < .. < i_k \leq n} P(A_{i1} \cap ... \cap A_{ik})$$

Appliquée à PROSEM, donnera :

Soit *m* un mot de la requête utilisateur et P_m la probabilité que le mot *m* soit dans le texte ou le corpus utilisé. Soit $RS(m)$ l'ensemble des mots liés à *m* par les relations sémantiques.

La probabilité que la requête contenant *m* donne un résultat peut être calculée en utilisant la formule de Poincaré comme suivant :

$$P(m) + P(\bigcup_{\substack{i=1 \\ mi \in RS(m)}}^{|RS(m)|} m_i) = P(m) + \sum_{\substack{i=1 \\ mi \in RS(m)}}^{|RS(m)|} P(m_i) + \sum_{k=2}^{|RS(m)|} (-1)^{k+1} \sum_{1 \leq i_1 < i_2 < .. < i_k \leq n} P(m_{i1} \cap ... \cap m_{ik})$$

Cette formule peut être améliorée dans le cas d'une « *requête sémantique* » composée de plusieurs mots représentant un sens particulier.

D'après ces calculs, PROSEM apporte sans doute une amélioration très importante en comparaison avec les méthodes classiques. Du fait que la probabilité qu'un mot *m* satisfasse une demande est toujours inférieure à la probabilité que *m* + *LS(m)* satisfassent la même requête.

Probabilité de la pertinence d'un mot par rapport à un contexte

La formule de calcul de probabilité d'apparition d'un mot dans un corpus ne tient pas compte de la notion de contexte, mais si l'on considère tout le processus de la démarche PROSEM, cette probabilité peut être corrigée par une fonction tenant compte du contexte. Ceci consiste à calculer, en plus de la probabilité d'apparition

d'un mot dans un corpus, la probabilité d'apparition du même mot dans l'ontologie du domaine et des contraintes (contexte). La formule de calcul de probabilités définie précédemment peut être utilisée à nouveau pour la prise en compte du contexte. Le produit des deux probabilités permet d'obtenir la probabilité globale correspondante à la probabilité d'apparition de mots (termes) *pertinents* dans le même document. La formule précédemment devient :

$P_{m:pertinent} = P(\bigcup_{\substack{i=1 \\ mi \in RS(m)}}^{|RS(m)|} m_i) * P(\bigcup_{\substack{i=1 \\ mi \in OD}}^{|OD|} m_i)$, où $RS(m)$ est l'ensemble des mots liés à m par les relations sémantiques et *(OD)* l'ontologie du domaine.

7. Conclusion et perspectives

Bien qu'ils existent des systèmes capables d'analyser et de traiter des contenus d'un point de vue sémantique, la relation qui lie le contenu à son utilisation est généralement peu ou pas du tout prise en compte. En effet, cette relation peut être d'une extrême complexité qui nécessite des approches et systèmes intelligents difficiles à mettre en œuvre capable de s'adapter en fonction du contexte d'utilisation. Dans cette article, nous avons proposé et montrer comment l'on peut utiliser PROSEM (PROjection SEMantique), une approche capable de gérer la sémantique dans son contexte, afin d'améliorer les performances de certains domaines d'application comme l'indexation, la recherche ou encore l'extraction d'information. Cela permet de répondre aux soucis relatifs à des énoncés ou à des mots auxquels peuvent correspondre plusieurs et différentes structures sémantiques en analysant fidèlement les relations sémantiques que peuvent entretenir les mots, les phrases, voire les expressions figées ou métaphoriques au sein d'un même texte et qui relève d'un même domaine, ainsi que les règles de raisonnement qui leur sont applicables.

Nous pouvons bien évidemment l'étendre à d'autres applications que nous étudierons dans nos travaux futurs, comme par exemple la fouille sémantique des données, la détection de données sensibles, la classification de documents, la génération des topics maps, la génération automatique de textes, etc.

Un autre aspect important qui va être traité consiste à affiner, à quantifier et à mesurer avec précision cette amélioration *(pertinence et optimisation)* dans des cas d'utilisation particuliers basés sur un corpus, une ontologie du domaine et des contraintes ainsi qu'une utilisation donnés.

Approfondir l'étude des relations sémantiques, déduire l'ensemble de toutes relations sémantiques potentielles qui puissent exister entre les mots puis spécifier les traitements à associer à chaque relation ou type de relations dans la démarche PROSEM sont d'autres aspects qui pourront être développés dans ce travail.

8. Bibliographie

The GATE platform: http://gate.ac.uk/

H.FADILI et M.CHAKIRI «Approche basée sur une « Projection sémantique » pour la compréhension automatique du texte : du mot au texte en passant par la locution», SITACAM, Agadir, 2009.

HERNANDEZ N., Ontologies de domaine pour la modélisation du contexte en Recherche d'information, Thèse de doctorat, Université Paul Sabaltier de Toulouse, 2005, 248 p.

BACHIMONT B., Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle, in Actes des journées francophones d'Ingénierie des Connaissances (IC'2001), Presse Universitaire de Grenoble, 2001.

AMARDEILH F., LAUBLET P. & MINEL J.-L., Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques, in Actes de la Conférence Ingénierie des Connaissances (IC'05), Nice, France, 2005, 12 p.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'2002), Philadelphia, July 2002.

Y. Li, K. Bontcheva and H. Cunningham. Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. Natural Language Engineering, 15(02), 241-271, 2009.

K. Bontcheva, V. Tablan, D. Maynard, H. Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. Natural Language Engineering. 10(3/4): 349-373. 2004.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.

Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

John Davies, C. Bussler, D. Fensel and R. Studer. The Semantic Web: Research and Applications. First European Semantic Web Symposium. Springer LNCS 3053. May 2004.

Sofia Pinto, Steffen Staab, York Sure, Christoph Tempich. OntoEdit Empowering SWAP: a Case Study in Supporting DIstributed, Loosely-Controlled and evolvInG Engineering of oNTologies (DILIGENT). C. Bussler and J. Davies and D. Fensel and R. Studer, First European Semantic Web Symposium, {ESWS 2004}, volume 3053 of LNCS. Springer, Heraklion, Crete, Greece. pp. 16-30. May 2004.

Fault detection system for Arabic language

Riadh BOUSLIM[1], Houda AMRAOUI[2]

[1] University FSJEG Jendouba Tunisia
bouslimi.riadh@hotmail.com

[2] University FSJEG Jendouba Tunisia
houda.amrawi@gmail.com

1. Introduction

The study of natural language, especially Arabic, and mechanisms for the implementation of automatic processing is a fascinating field of study, with various potential applications. The importance of tools for natural language processing is materialized by the need to have applications that can effectively treat the vast mass of information available nowadays on electronic forms. Among these tools, mainly driven by the necessity of a fast writing in alignment to the actual daily life speed, our interest is on the writing auditors.

The morphological and syntactic properties of Arabic make it a difficult language to master, and explain the lack in the processing tools for that language. Among these properties, we can mention: the complex structure of the Arabic word, the agglutinative nature, lack of vocalization, the segmentation of the text, the linguistic richness, etc.

In that perspective, our project aims to develop a system to detect errors in spelling, structure and conjugation of the Arabic language. In this article we will proceed as follows. In the first section we'll present some approaches used for the correction of errors. The second section will be devoted to detailed studies of our proposed system. In the last section, we'll perform experimental tests to evaluate the performance of our system.

2. State of the art

2.1. MASPAR

A multi-agent system is a system of agents' group that communicate with one another to provide answers about a goal to achieve.

MASPAR is a system of analysis of Arabic texts based on the approach of multi-agents. It consists of a set of agents, using a direct communication by sending messages. These agents work together in order to make syntaxes' analysis of a sentence given by the user by determining its syntax composition. (tree, je ne sais pas si ca existe!!!c un mot relativement technique, il faut voir...)

2.1.1. MASPAR System Limits

The major drawback of such system is the time taken by the agents for communication and interaction.

One might also note that the MASPAR system does not detect errors of conjugation. Also, it has a non-ergonomic interface.

3. Proposed System

3.1. General Description

Our system (Figure 1) is designed to detect errors in spelling, structure and conjugation in a non- vowelized Arabic text. It consists of five phases, each uses the information received from the previous phase to finally get a text containing the least number of mistakes.

The segmentation phase consists on dividing the text into sentences and then into words. The lexical phase subsequently receives the word and checks its existence in the database of words.

After verifying that this word belongs to the language, the phase labelling associates the word it has received the possible morph syntactic labels, this makes the word ambiguous, hence the need to remove this ambiguity by passing phase disambiguation, which in applying certain rules, is used to assign to this word the most suitable label.

To correct the word "Wc", we must compare it with the database of words that we have, if this word belongs to our dictionary, it means that "Wc" is a correct word otherwise our system will detect a misspelling.

The algorithm then verify the proper structure of this sentence, otherwise the system will detect a structure fault. Finally, our system is also capable to detect the faults of conjugation in a sentence.

We, first, introduce the general architecture of our system.

Figure 1 : Proposed system

3.2. Detailed Description

When receiving an electronic text to analyze, our system launches the first phase which is segmentation. This phase begins with the identification of the text' sentences based on punctuation signs then on the words in each sentence. Subsequently, the words in each sentence will be transferred one by one to the lexical phase.

This will verify whether the word belongs to the language or not by checking its existence in our database of words. Subsequently, the word is sent to the next phase. The phase label is responsible for providing possible morph syntactic characteristics of each received word (from the lexical phase). This means that a word can't go to the labelling phase unless its belonging to our database has been confirmed within the lexical phase.

Because each word can have several labels, the analysis of the word can face certain ambiguity. That's why we must use rules to reduce this ambiguity. Therefore, disambiguation phase is triggered to limit the number of labels associated with the word and assign a single label at a time.

Once the ambiguity is removed, we get into the final phase of the system which role is to apply rules that enable to compare the analyzed structures. This helps detect errors in structure and conjugation.

```
ALGORITHM Editor
STARTERS: Wc: the word of the sentence
          Phrase: the input sentence
          BaseXml: the database contains dictionary words
          BaseReglesStruc: the database of structural rules
          BaseReglesConjug: the database according to the rules of conjugation
START
      FOR each Phrase DO
        FOR each Wc of Phrase DO
          If (Wc, BaseXml) == false then
            Write (Wc 'is incorrect')
            Otherwise
            Type□ ReccupererType (Wc, BaseXMl)
              Structurephrase □ Type
          End if
        End For
        Compare (Structurephrase, BaseReglesStruc)
```

```
                If (compare == true) then
                    Write ('the structure of the sentence is not correct')
                  Otherwise
                  If the structure contains a verb then
                        Apply (BaseReglesConjug, Structurephrase)
                  End if
                 If (Apply==false) then
                     Write ('the combination is not correct')
                 End if
              End if
          End For
END.
```

3.2.1. Segmentation

This phase consists on dividing the text into sentences and the sentences into words based on markers at the beginning and the end, for example points, semicolons, colons…

3.2.2. Lexical Analysis

This phase checks the belonging of each word to the language, obtained from the segmentation phase based on the data base of the words available.

Verify the existence of the base in the lexicon : We must ensure that the words introduced constitute the basics of the Arabic language. For that reason, we verify the existence of the base in the lexicon. We have to consult the database of Arabic words, if the extracted base coincides with a word from the database; we conclude that the word exists in Arabic.

3.2.3. Labelling

This operation aims to add to the words linguistic information with morphological or syntactic nature in order to identify them.

We have presented several possible tags of the word minimum (prefix + base + suffix):

However, the lack of vocalization does not accurately determine the proper etiquette of the word which causes a certain ambiguity. To reduce this ambiguity, we will proceed to the next step.

3.2.4. Disambiguation

A disambiguation is needed to limit the number of labels of these words and subsequently improve the detection of grammatical errors.

Compatibility Rules: It can reduce the ambiguity of a word by associating it with one type at a time, so the sentence containing the ambiguous word has more than a structure based on the number of labels that word. Subsequently, the system associates to the word the suitable type according to the structural rules.

3.2.5. Detecting faults

For the detection of faults, we can use rules of grammar. These rules describe correct grammatical patterns. For this, we have defined a basic structure rules and another different basis for the conjugation rules. If some text does not match any rule, a structural or conjugal error is detected. To detect structural faults, we'll compare our sentences' structure with the basic structural rules, if this structure does not coincide with any rule, then a lack of structure will be detected, otherwise, if the structure is correct and if it contains a verb, since the combination only applies to the verb, our system will have access to our database conjugation, satisfies a certain compatibility between pre-and post-basic core that typically accompany the verb, if our sentence presents a bad combination when a fault is detected. In the end, the user receives a text containing errors detected with staining of these faults, each depending on the type of errors detected.

3.2.6.The databases used in the system

```
<MOTS>
  <Noms>
    <NomsPropres>
      <NomsPropresFeminins>
          <NomPropreFeminin>أسماء</NomPropreFeminin>
          <NomPropreFeminin>أمل</NomPropreFeminin>
          <NomPropreFeminin>إيمان</NomPropreFeminin>
        </NomsPropresFeminins>
      <NomsPropresMasculins>
          <NomPropreMasculin>أكرم</NomPropreMasculin>
          <NomPropreMasculin>إلياس</NomPropreMasculin
          <NomPropreMasculin>أيمن</NomPropreMasculin>
        </NomsPropresMasculins>
      </NomsPropres>
    <NomsPluriels>
```

XML database for the detection of spelling errors

To detect if the spelling of a given word is correct or not, the verification process run through the XML tree of the dictionary and compare the word with the word list file. It sets out below a portion of our base words.

XML Data Base for the detection of structural faults

After ascertaining that the specified words are spelled correctly, we assess at this level whether the sentence structure is coherent or not by comparing it with the base of the structures we have created an XML file with the following form:

```xml
<ReglesApplicables>
  <ReglesPhrasesVerbales>
     <regle>verbe NomPropreFeminin</regle>
     <regle>verbe NomPropreMasculin</regle>
     <regle>verbe NomPluriel </regle>
  </ReglesPhrasesVerbales>
<ReglesPhrasesNominales>
   <regle>NomPropreFeminin verbe </regle>
   <regle>NomPropreMasculin verbe </regle>
```

XML Data Base for detecting faults conjugation

Our system can also detect conjugation errors. To handle this, we used an XML file as follows:

```xml
<PronomPersonnel valeur="أنتم">
  <PresentSimple>
     <prebase>ت</prebase>
     <PostBase>ون</PostBase>
   </PresentSimple>
  <PresentNegation>
     <prebase>ت</prebase>
     <PostBase>وا</PostBase>
   </PresentNegation>
  </PronomPersonnel>
```

4. Test and Validation

We choose to assess the performance criteria that are available: the ergonomics and the response time chosen by our system.

Regarding ergonomics, performance analyzers must have a user-friendly interface, presenting a number of functionality to help users better handle this interface to manage the features offered by the system.

The speed of response is another important constraint for parsers for, to be useful in the real world, they must return a response very quickly.

4.1. Experiments

Our experiments on the system relate texts of Arabic literature in various fields. We introduced those relating to the field of Medicine, Marketing, Economics and Arabic grammar.

(-): If no fault is detected.

(+): If an error is detected.

Sentences	*Detection of spelling errors*	*Detection of structural errors*	*Detecting of conjugation errors*
يبحث في أصول تكوين الجملة و قواعد الإعراب	(-)	(-)	(-)
و يبحث فى فى أصول تكوين الجمة و قواعد	(+)	(+) (+)	(-)
يتكون جسم الإنسان من أجهزة مختلفة الوظائف	(-)	(-)	(-)
التسويق هو مجموعة من العمليات أولأنشطة	(+)	(-)	(-)
التسويق هو مجوعة من العمليات أو الأنشطة، تشبعوا رغبات العملاء	(+)	(-)	(+)
أنتم لم تذهبون	(-)	(-)	(+)
أنتم لم تذهبوا	(-)	(-)	(-)
يأخذ إيمان أقراص	(-)	(-)	(+)
يأخذ أيمن أقراص	(-)	(-)	(-)
تذهبين إيمان	(-)	(-)	(+)
تذهب إيمان	(-)	(-)	(-)
هما لن يذهبان	(-)	(-)	(+)
لم يكتبوا الجملة	(-)	(-)	(+)

هم لم يكتبوا الجملة	(-)	(-)	(-)
أيمن يذهب	(-)	(-)	(+)
النحو العربي هو علم، تبحث في أصول تكوين الجملة و قواعد الإعراب	(-)	(-)	(-)
يذكر أن في مثل ذلك المكان	(-)	(+)	(-)

To evaluate the error detection, we use the rate of accuracy (standard indicator classification [4]). This indicator is between 0 and 1.One being the perfect result.

To calculate this index, we needed to appoint different sets.

Let D be the total set of words, incorrect words D + and D-words correct. D + and D-form a partition of D. Let R be the set of words identified as erroneous. Some words are part of R + D D-other.

The precision is out as an index of the proportion of words identified as erroneous. Its formula is:

Detection Accuracy = | D + ∩ R | / | R |

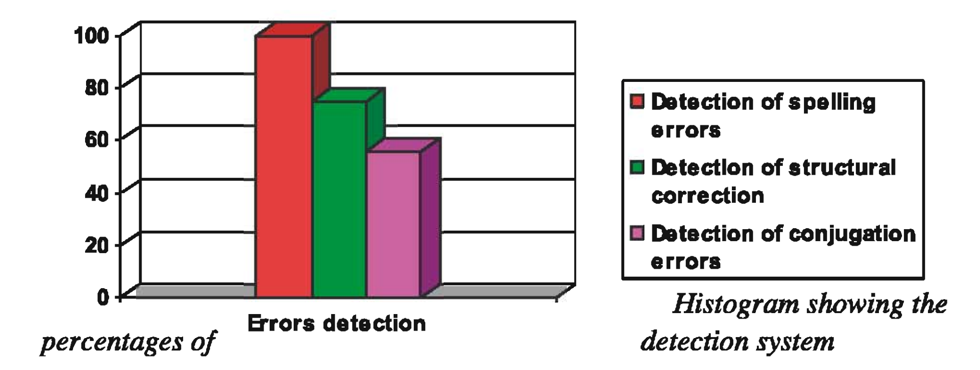
4.2. Results and interpretations:

We note that our system has a good detection for errors in spelling and structure. (Indicator precision = 1 for the detection of spelling errors and 0.75 for structures). Indeed, we get a quick response if the word entered is incorrect or the structure is wrong. We therefore have a very high proportion of errors actually detected. We can also note the good accuracy of the wrong word or structure which facilitates the coloration of errors.

We can also note that our system has a medium detection for errors in conjugation (Indicator accuracy = 0.56). This is for several reasons:

Note first the difficulty of the Arabic language in particular as regards to the conjugation.

Then, our system took only where the verb is conjugated in the simple present and present Negation and verifies the compatibility between pre-and post-foundation bases with different personal pronouns. By cons, although the average conjugation fault detection, our system provides a new paradigm as it has treated the most difficult in the Arabic language is the conjugation.



Histogram showing the percentages of detection system

5. Conclusion and Perspectives

The information retrieval and text mining in Arabic is a major challenge. We are interested in this work to develop an application to detect errors in spelling, structure and conjugation in the Arabic text.

The development of this project allowed us to familiarize ourselves with the Java language, a language in the promising field of programming technologies. It allowed us to consolidate our knowledge on various techniques including manipulation of XML.

The work we have done is a response to the objectives set at the outset of the project. However, it can evolve by considering several extension elements.

We can consider adding propositions for the wrong words in order to improve the performance of our system. We can also add more functionality to our tool such as translating the input text from one language to another following the user's choice. We can also handle the case semantics and the texts vowels.

6. Bibliographie

Aloulou C., Belguith H., Hadj Kacem A., and Hamami M. (2005). *The Book System implementation*

on a MASPAR approach MULTI-AGENT, Faculty of Economics and Management of Sfax.

Wilson W., Wei L. and Mohammed B. (2006). *Scoring for Integrated Spelling Error Correction,*

Abbreviation expansion and Case Restoration in Dirty Text. School of Computer Science and

Engineering Software University of Western Australia

Douzidia Fuad S. (2004). *Automatic summarization Arabic*. University of Montreal: Department of

Computer Science and Operations Research Faculty of Arts and Sciences

Guillaume P. (2005) *Fixed spelling in context*. Compiegne University of Technology.

Exercises in Arabic Indexing : Finding Repetitions in the Quran

Khalil Honsali[1], Mohammed Majid Himmi[1], El houssine Bouyakhf[1]

[1]LIMIARF/FSR, U. of Mohammed V Agdal, Maroc

{k.honsali}@gmail.com

1. Introduction

Experience acquired within the Lase project [Tazzit et Al. 2009, Sabri et Al. 2006] has led us to develop and test indexing strategies for different types of Arabic corpora [Maamouri et Al. 2004]. In this paper, we report our use of indexing techniques to perform batch statistical analysis of the Quran, namely to count occurrences of a word or word phrases. Although this problem may appear simple, it hides underneath several computing challenges. We present our results as a gift for God lovers.

2. Indexing the Quranic Corpus

2.1. Quranic Corpus

We used the Quranic corpus used by most sites, and which is available at [quran.com]. There

are different formats available. We chose plain text with simplified script (no tashkeel), a

version of the text with minimal pre-processing overhead, namely no tashkeel and other punctuation elements. Each line of the text contains a chapter/verse combinations in the

following format:

ميحرلا نمحرلا للها مسب|1|1

ChapterId | verseId | verseText

The file is parsed and the whole Quranic corpus is loaded into memory in few milliseconds,

as a Quran object, containing a list of Verse object, each verse has meta information plus a list of word objects, each Word has the text and a list of Location objects. A Location object is all

of the word offset in the Verse and its order. This is the object used extensively for search, since it indicates the location of a word and is used to identify repetitions.

After the file is parsed, the Quran index is constructed. This is explained in the section below

2.2. Inverted Index Construction

To build the index, the program must loop over all the verses/chapters, and extract their words.

When a word is encountered for the first time, it is added to the index; if it already exists there, then a new location is added to its list. The algorithm below explains this process.

For each *verse* V

parse *word* list -> list(W)

For each *word* W

If *INDEX* contains W is **false**

add W and W.*location* to *Index*

Else

fetch W in *INDEX*

add new *location* to W

Figure 1: Algorithm for indexing the quran

It is important to note that the index should be clean from frequently used grammatical

particles (و، إلى، نم، ،،،) , and other words that are meaningless and constitute noise in processing, also known as stopwords. Moreover, other indexing processes such as tokenization and lemmatization to extra word roots are usually necessary however not in this context where we consider each lemma being unique.

3. Counting Repetitions

3.1. Single word repetitions

Once the index is constructed, information about the different locations of each word is

available and that corresponds to the number of repetitions. The more locations a word has the more it is repeated in the text.

3.2. Multi-word repetitions

For finding word phrase repetition, more thorough index analysis is necessary. We undertake

a straightforward approach in counting phrase repetitions by following the algorithm below. First we discard non-repeated single words, and hence construct a list of, repeated-words, which are words with more than one location (minus stopwords). For each word in the list, and for each of its locations, we search for words with locations directly following the current location.

```
For each word W in list of repeated words
Fetch location list for W-> locationList        //fetch repetitions
For each location Loc in locationList           //for each repetition
        nextLocation = W.loc + W.length + 1;                  // determine next location
        nextWord = getWordAtLocation( nextLocation );     // ..and next word
        Fetch location list for nextWord -> locationList2 //fetch all locations for that
word
        For each location Loc2 in locationList
                If( loc2 == loc )
                Results.add( loc2, nextWord);
```

Figure 2: Algorithm for finding multi-word repetitions

4. Experiment

The table below describes the different processing steps and performance costs associated

with them. The machine is a Centrino dual core processor (1.6G) and (2G) of RAM. Parsing file, building index, search 1 word repetitions, search multi-word repetitions

And here are some results for the data as follows:

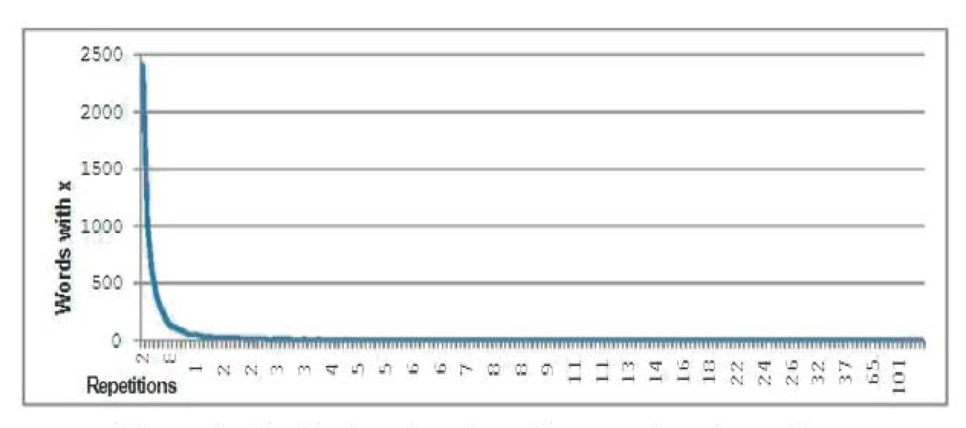


Figure 3: Graph of number of words vs. number of repetitions

The table below shows some of the results found:

	Most repeated word	*Number of repetitions*	*Total number of repeated words*	*Processing time*
Single Word	الها	*2265*	*1600*	*< 1s*
2-word phrase	*NA*	*NA*	*NA*	*4 - 5 mins*

5. Related Work

Concerning Arabic Indexing, there exists some work related to Arabic morphology analysis

[Sawalha and Atwell], or how to extract the roots from Arabic words. As far as the Quran is concerned, an interesting project done by [Dukes and Hasbah 2010], consists of constructing an ontology of the concepts in the Quran, as well as a grammatical tree analysis. There isn't much work readily available on counting repetitions of the Quran [Ali 1427], but it does not explain the technical details behind it.

6. Conclusion

We have successfully implemented an index and a program for searching repetitions within

an Arabic text. The method has been tested on the Quranic corpus and interesting results were shown. However, currently tashkeel is not supported and does have an impact on making distinctions between words.

7. Future Work

We are in the processing of generalizing this algorithm to support an infinit number of words in a phrase; consequently, the processing time should be exponentially higher. Hence, it is also considered to distribute the algorithm using the latest technology known as map/reduce.

8. References

Naïma Tazzit, Abdellah Yousfi, El Houssine Bouyakhf (2009) "Design and Implementation of an

Information Retrieval System by Integrating Semantic Knowledge in the Indexing Phase" ICGST-

AIML Journal, ISSN: 1687-4846, Volume 9, Issue I, February 2009

Souad Sabri, Abdellah Yousfi , El Houssine. Bouyakhf. (2006) "Un système d" analyse morphologique des noms dérivatifs arabes". JETALA" 2006, Rabat, 05-07 juin 2006, Maroc.

Ahmed Abdel-Fattah M. Ali (1427) Word Repetition in the Quran - Translating Form or Meaning? J. King Saud Univ., Vol. 19, Lang. & Transl., pp. 17-34

Mohamed Maamouri, Ann Bies and Tim Buckwalter (2004). The Penn Arabic treebank: Building a

large-scale annotated Arabic corpus. In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Kais Dukes and Nizar Habash. Morphological Annotation of Quranic Arabic. Submitted to the seventh

international conference on Language Resources and Evaluation (LREC-2010). Valletta, Malta, 2010

Majdi Sawalha, Eric Atwell. (2008) "Comparative Evaluation fo Arabic Morphological Analyzers".Coling Companion Volume. Posters and Demonstrations, pages 107-110.

A Tool for Annotating Texts with Morphological and Syntactic Information

Violetta Cavalli-Sforza, Hasnae Rehioui, Leila Bahri

Al Akhawayn University, Ifrane, Morocco
V.CavalliSforza@aui.ma, H.Rehioui@aui.ma, L.Bahri@aui.ma

1. Background and Motivation

We describe an Annotation Tool for adding morphological and syntactic annotations to texts. The Annotation Tool is being developed in the context of a larger application: the Arabic Reading Assistant or ARA for short (Cavalli-Sforza & Chekayri, 2010). The ARA system is targeted at helping students of Modern Standard Arabic (MSA) read increasingly complex texts. The texts' vocabulary and grammar content reinforce already acquired language concepts and challenge the learner with new ones, leading her along a predefined language learning curriculum (broadly based on Brustad, Al-Batal., & Al-Tonsi, 2004, 2006, 2007). While taking into consideration her apparent level of mastery. The Annotation Tool is used to annotate the texts with the specific lexical, morphological and grammatical concepts that are targeted by the curriculum. While our objective in developing the Annotation Tool has been specifically to support our language learning application, by augmenting texts with annotations relevant to selecting passages for testing and tracking student learning, the tool itself is designed to be easily repurposed for other applications and languages.

The version of the tool described in this paper represents an intermediate version intended for internal use, and still suffers from some design and development faults. We are currently revising the tool to improve its ease of use and increase its functionality. In the remainder of this paper, we will refer to the tool itself as the Annotation Tool or AT, reserving the name 'annotator' for the user of the tool, the human annotator.

2. What is an Annotation?

Most generally, an annotation is a group of one or more feature values that may be associated with a text fragment – a single word or a multi-word segment of the text. In our system, we distinguish between

- *morphological features*, which pertain to individual words, even though those words may be compounds including various prefixes and suffixes; and
- *grammatical features*, which provide information about multi-word fragments in the text.

Jointly the two are referred to as *morphosyntactic features*. The distinction between the two, at the level of the AT, is not a strong one. It mostly affects which menu the features and their values are chosen from, but does not substantially impact the way in which they are handled the AT or the annotator.
An example of a morphological feature is the part of speech of a word or suffix (e.g. for personal pronouns suffixes) with values such as *Noun*, *Verb*, and so on, or a feature such as *Gender*, with values *Feminine* or *Masculine*. An example of a grammatical feature is the construction *iDafa* (إضافة), which would be associated with a fragment at least two words long and possibly longer.
While the most basic annotation associates an individual feature value with a text fragment, some annotations are complex annotations and include a collection of feature values. These collections, called Grouped Annotations, can be composed via the interface, given a name and saved for future reuse. For example, the annotator may find it convenient to have a complex annotation that includes the features values *Noun*, *Human*, *Feminine*, *Plural*, *Sound*. This group can be created once, and then named, saved, and reused. It can also serve as a basis for creating more specific and complex annotations, such as the *Nominative* and *Genitive-or-Accusative* variations of the above.
When associating annotations with a text fragment, only the feature value is stored. The feature name is implicit and values should be chosen to be unique and, possibly, mnemonic. Should this constraint prove to be difficult to abide by, it would not be difficult to change the system to store the feature name as well and thus create a 'namespace' for feature values to tolerate duplicate values with different meanings in different feature namespaces.
The AT stores annotations in an external file that is separate from the text to which the annotations refer. The file is in the same directory and has the same name as the

text file but uses a different extension.[1] The external file is readable via a text editor, but should not be manually modified. For example, in the current format for external files, the annotation line "ip*4*264:يعود" associates the annotation *Present* (internally represented as ip) with the text fragment (word) يعود starting at position 264 and of length 4. Each line in the annotations file contains all the feature values associated with a specific fragment. Annotations can also be nested within each other. For example, the annotation "iDa:9:0:ابن بطوطة" says that the fragment "ابن بطوطة", starting at character 0 of the file and of length 9 is an *iDafa*. Associated with the same fragment, there may be annotations for individual words, which would have their own entry in the file. For example, for the word 'ابن' one might find "s, m,NOUN:3:0:ابن", indicating that this word is a *Noun* and is masculine and singular.
The source of the annotations themselves are two feature-value 'trees', one for morphological features and one for grammar features, that are created via a special interface provided by the system and described in greater detail below.

3. Displaying Annotations

Figure 1 shows the main window of the Annotation Tool. Below the menu and some general file information, the right uppermost pane displays the text itself. The annotations a in the text are displayed in two different ways. To the left of the text pane is line-by-line listing of annotations associated with each fragment, word or phrase; each feature is displayed on a separate line. By default, the display is ordered by starting position, which causes annotations pertaining to the same fragment to be positioned in close proximity to each other.

[1] In the future, it may be desirable to relax this assumption and allow the annotation file to have a different name, so that the same text can be annotated differently. However, at present, this is does not appear necessary.

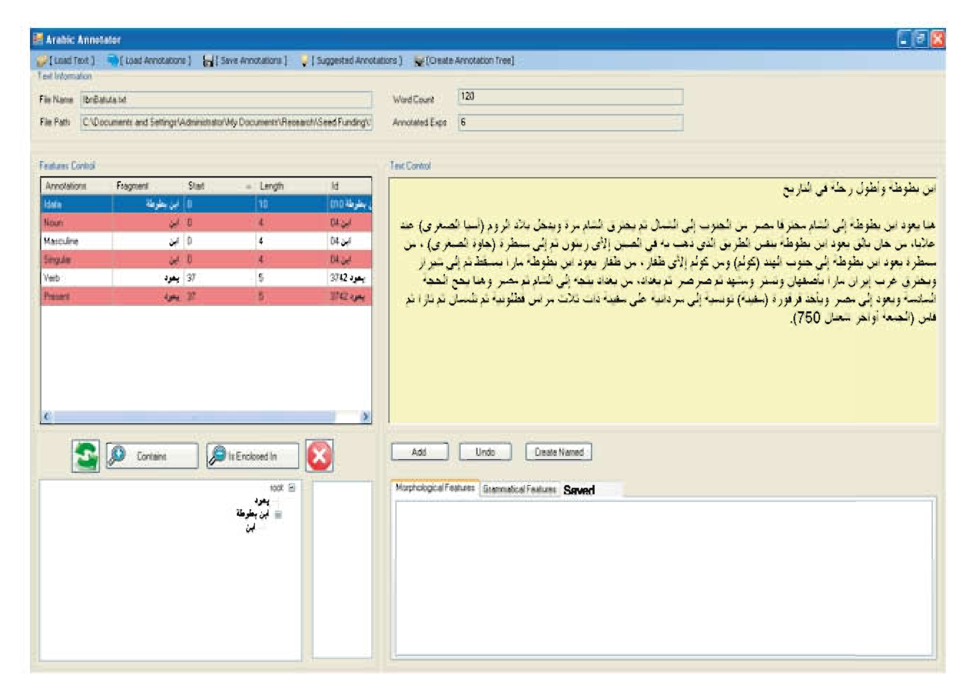


Figure 1: The basic Annotation Tool interface showing a text with nested

Below the line listing of features, a tree view, also sorted by position of the fragment in the text, shows the nesting of annotated fragments within each other. The root of the tree is the entire text. Contained annotations are shown as the children of the containing ones if the containing nodes are expanded. The tree provides a high-level view of all the annotations in the text and can be used to navigate through them more quickly than by scrolling through the single feature display of annotations. The **Contains** button located between the two panes, if clicked when a fragment containing other annotated fragments is selected, causes the feature display to focus on just the contained fragments (in Figure 1, only the features for 'ابـن' would be displayed). Clicking on the **Is Enclosed In** button to its side performs the opposite operation: given a selection that represents an annotated fragment embedded in another, it switches the display to focus on the containing fragment. The button refreshes the display, returning it to the original one-feature-per-line display. Finally, the button is used to delete one or more feature annotations selected from the one feature per line display.

Clicking on any of the tree nodes changes the highlight in the text area to point to that word, as does clicking on a feature line in the detailed feature display. In the future, the three panes (tree view, feature view and text) will be completely

coordinated so that clicking in any one of the three will cause the other two to (re)position the highlight on the selected fragment.

4. Adding Annotations

The AT supports adding annotations to texts in two different ways: manually and by choosing among annotations suggested by an external program. Figure 2 shows a simple example of adding a manual annotation. To add an annotation manually, the annotator selects from the feature hierarchy shown below the text and applies it to the desired text fragment by using the Add button. More details are provided below about the construction of the feature hierarchy.

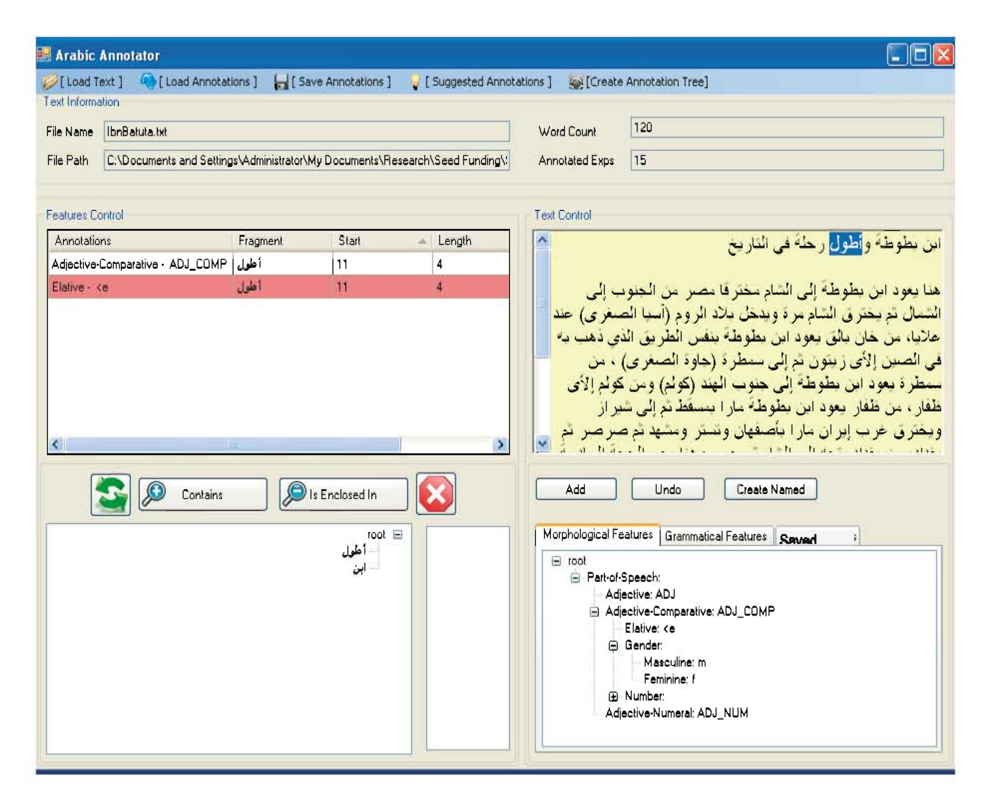


Figure 2. Adding annotations manually to the text with the feature hierarchy

The annotator can select an annotation and apply it repeatedly to several text fragments, or select a fragment of the text and apply different annotations to it. In addition, as mentioned earlier, it is possible to select a combination of annotations, assign a name to the combination and save it so it can be reused elsewhere. The named combinations are stored with the AT's resources, not with the texts, and are

explicitly loaded into the AT, as are the tree-structured feature menus. The tab named Saved Combos gives access to those named combinations. Using a named combination allows multiple feature values to be associated with a fragment with a single Apply action.

The feature hierarchy approach, which is used for morphological and grammatical features, although using different hierarchies, helps the user annotate correctly and reasonably quickly. The designer of the feature hierarchy, who must have good linguistic knowledge, insures the hierarchy's correctness by organizing features and their values in such a way that the descendants of a node only include features relevant to the ancestor nodes. For example, the descendants of a *Part-of-Speech* = *Noun* node do not contain a *Tense/Aspect* feature, but do include several features appropriate for nouns such as *Case*, *Number*, and *Gender*. Similarly, a *Case* feature should not be associated with the descendants of a *Verb* node. While a well-designed feature-hierarchy insures that the annotator does not choose features and values that are inappropriate for the category of the word or fragment, it is still up to the annotator to choose the correct analysis for the fragment to be annotated.

The feature hierarchy is represented in what might be called, for lack of a better name, an OR-XOR tree, to distinguish between combinable and mutually exclusive features. An initial graphical interface for building the tree is shown in Figure 3. Each node of the tree represents a feature name or a feature value. The tree contains three types of nodes, XOR, OR and mandatory, though their difference is not displayed visually in the current interface.

The *Part-of-speech* node, representing the name of a feature, is itself an OR (combinable) node and has as its children a collection of XOR nodes that provide different mutually exclusive options for the value of this feature (*Adjective*, *Adjective-Comparative*, *Adjective-Numeral*, etc.). The associated tags shown after the semicolon (e.g., ADJ_COMP) are the internal tags used for the different values and are currently displayed for debugging ease.

Under the *Adjective-Comparative* node, there are two types of nodes. The *Gender* feature node is an OR node: it can be combined with other OR features (for example, with *Case*). The two children of *Gender* are XOR nodes, since a noun cannot be masculine and feminine at the same time.[2] Making the values of features into mutually exclusive siblings (XOR nodes) precludes selecting combinations of

---

[2] In a few cases it may be desirable to label something as having multiple values of the same feature. For example, some nouns can be considered as either masculine or feminine. In this case a special feature value is used that includes both masculine and feminine and is mutually exclusive with just masculine and just feminine.

features that are inconsistent (for example, choosing the part of speech for a word to be both a *Noun* and a *Verb*, or attributing to a *Noun* both a singular and a dual value for the Number feature). In contrast, any number of combinable sibling nodes can be selected. When adding annotations, selecting an XOR node deselects any other XOR node that is its sibling, whereas multiple sibling OR nodes can be selected at the same time.

A third type of node present in Figure 3 is the mandatory node, of which the *Elative* node is an example. It is used for a feature value (<e) that is always associated with elative (comparative or superlative) forms of adjectives and must be present if its parent is added as an annotation.

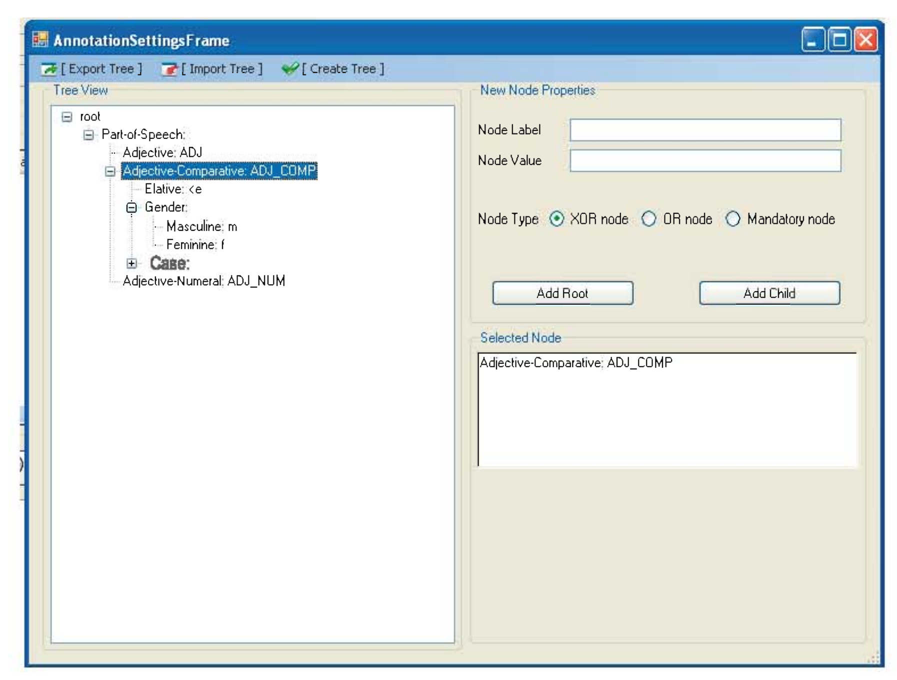


Figure 3. Initial version of interface for building the feature hierarchy

Some of the nodes in the OR-XOR tree (e.g., the *Adjective-Comparative*, *Elative*, *Masculine*, and Feminine nodes), including all leaf nodes, have an associated value that are added as annotations to the text fragment. Those that do not have values are used to group the nodes below them. The different types of nodes have a different behavior when used in adding annotations. Selecting a node that has an associated feature value sets it up for having that value added as an annotation to a text fragment when the annotation is applied, whereas selecting a node that has no feature value associated with it merely expands the node to give access to its children.

The feature hierarchy is one of the ways in which the AT helps speed up annotating a text. Selecting a node in the tree and adding it as an annotation automatically adds all the feature values associated with nodes between it and the root. Feature values associated with mandatory nodes are also added if their parent is. Since combinable nodes will have feature values in common because of common ancestry, the process of adding annotations checks for duplicate feature values and discards them.

The Annotation Tool speeds up the annotation process in other ways. On the one hand, the AT does not force the annotator to annotate every word or phrase in a text, nor does it require that fragments that are annotated be annotated with all the relevant features. Using the feature hierarchy the annotator can choose features that are as deep as the leaves or that partway down the hierarchy. In both cases all feature values between the selected node and the root will be added, together with feature values associated with any mandatory node. For example, in Figure 2 the annotator selects and adds the *Elative* feature, which automatically adds its parent, *Adjective-Comparative*. Selecting *Adjective-Comparative* would have had the same effect, since *Elative* is a mandatory feature. The annotator is not forced to select and add a *Gender* feature value, *Masculine* or *Feminine*; however, should she choose to do so, the selected *Gender* value, as well as *Elative* and *Adjective-Comparative* would be added too. Similarly, the annotator may choose to add *Gender* annotations but not add *Case* annotations. This flexibility is important in our application, since the purpose of annotation is to attach to texts information that can be used to retrieve passages using criteria that are appropriate at specific points in the curriculum. Texts can be annotated to contain just those features that are of interest at the time; other features can be added at a later time if it becomes desirable to do so for the purpose of making the texts usable at different points in the curriculum.

Finally, the manual annotation functionality, while provided for all fragments, is actually most useful as a way of adding syntactic annotations to multi-word fragments of the text or partial morphological annotations to specific words in the text. It is assumed that, for complete word-specific morphological annotations, rather than individually selecting annotations for each word, the annotator will be selecting from annotations suggested by an external morphological analyzer (in our case the SAMA analyzer, a descendant of the Buckwalter morphological analyzer). To allow the annotations to be selected among proposed ones, the AT reads in a file of possible analyses for each word in the text, if necessary performs a translation between the annotations suggested by the external program and those used by the AT using a predefined mapping, and presents to the annotator all the suggested analyses for each word at a time. The annotator's choice is then stored along with annotations added manually to other parts of the text.

5. Tool History, Current State and Future Work

The Annotation Tool described above is currently undergoing its third revision. The first version was developed in Java by a student as a capstone project for her bachelor's degree in Computer Science. Her work provided a platform for experimenting with our initial ideas for adding both morphological and syntactic annotations to potentially long texts. The tool had two distinct annotation 'modes' that the user could choose between: 1) an *annotation-driven* mode, in which a specific annotation, consisting of a feature or set of features, could be selected and applied to all the suitable text word or phrases; and 2) a *text-driven* mode, in which the user would choose the text fragment first and then apply the desired annotation. It was also partially integrated with a database storing the text annotations and allowed searching for texts in the database that met different criteria in terms of vocabulary and morpho-syntactic content.

The second version of the tool, and the one described in this text, was rewritten in C#, in order to have better support for Arabic text operations. It was mostly the work of another undergraduate student, Hasnae Rehioui, with a small amount of additional work done by Leila Bahri, a Master's student, to clean up the display of individual annotation features and give a means of including the morphological and grammatical feature trees into the AT. It also provided a more sophisticated way of displaying and handling of annotations, and integrated the ability to select from annotations suggested by an external morphological analyzer. Its use also showed that there was no real need for the interface to separate the two annotation modes: both fell out quite naturally from the interface depending on what the user chose first, the text or the annotation.

The third revision of the AT, currently underway, aims to retain the above functionality while making the interface clearer, less cluttered, and generally more usable and powerful. In the first place, since a text may have extensive annotation information associated with it, it is crucial that the display of information be flexible and as easy to navigate as possible. Several or all of the words and many phrasal fragments may have annotations and the annotations can include several features. The annotator must have access to the individual features that are part of a more complex annotation, as well as see easily which text fragments are annotated and how they relate to each other. The Annotation Tree display gives a high-level view of how different fragments of the text containing annotations relate to each other, but the features display can be long and hard to navigate.

The newest version of the AT will keep displaying the individual features on separate lines but use a two-level display. The *overview display* will include only one entry for each annotated fragment, including the same information as before (the fragment itself, start, length, and one feature, in order of starting position by default. If more features are present, their presence will be signaled by a symbol indicating the possibility of expanding to view the full list of features. If a phrase and its individual words contain annotations, they will occur in the proximity of each other in the display. In the *expanded display*, all items will be shown with all their features. This display will allow reordering the annotations by feature (value) and, within the feature group, by other criteria. This expanded display can be used to check for consistency of feature attribution across the text and to perform additions or deletions of features on entire groups of entries. The interface for suggested annotations is also being redesigned to support a similar style of display and, generally speaking, to facilitate the interaction between manual annotation and selection of suggested annotations.

Another change that is currently underway is the database back end in which the annotations will eventually be stored. Currently the AT stores annotations in an external text file whose contents are to be loaded into an SQL database for use with the ARA environment. In the future, it will be possible to both right to a file and to insert the annotations directly into the database. So far, we have been working with an SQL Server 2005 database running on a university server. To offer more flexibility and independence to developers and users, we have decided to also support MySQL and to allow the database to be running on the same machine as the annotator in addition to on a remote server.

Finally, we note that we aim to make the AT usable for annotating text resources in other languages and to be able to interface with other systems. The latter is facilitated by the use of external annotation text files whose format is not language dependent. The AT user interface itself is intended to be bilingual and could be easily extended to be multilingual and/or work with a different language pair. Our immediate use of the AT is for annotating MSA texts and therefore, in the near future, the AT will 'speak' English and MSA, easily toggling between the two. Considering also that the tool is currently being used by a mixture of English and Arabic speakers, but that the ultimate users of the tool may be primarily Arabic speakers and may be more comfortable with the terminology traditionally used to discuss Arabic grammar rather than English grammatical terminology, the annotation tree will include both English and Arabic labels, permitting the display of either or both. The interface for incorporating the morphological and

grammatical annotation trees minimally constrains the form and content of the annotations that can be handled. Therefore we hope, when (re)development is completed, to share it with other colleagues who might find such a tool useful for their applications and we welcome input from colleagues working on similar tools or using similar tools who could have an interest in using the Annotation Tool once we are ready to release it.

5. Acknowledgments

We gratefully acknowledge the financial support of Al Akhawayn University for the Arabic Reading Assistant project, and the work done on the first version of the AT by Yousra Aafer, the first student at AUI to work on the ARA project. Her efforts won Stonehenge Award, given yearly to the best capstone project by Al Akhawayn University Alumni Association.

References

Brustad, K., Al-Batal M., & Al-Tonsi, A. (2004), *Al-Kitaab fii Ta'allum al-'Arabiyya with DVDs: A Textbook for Beginning Arabic*, Part One, Second Edition, Georgetown University Press.

Brustad, K., Al-Batal M., & Al-Tonsi, A. (2006), *Al-Kitaab fii Ta'allum al-'Arabiyya with DVDs: A Textbook for Beginning Arabic*, Part Two, Second Edition, Georgetown University Press.

Brustad, K., Al-Batal M., & Al-Tonsi, A. (2007), *Al-Kitaab fii Ta'allum al-'Arabiyya with DVD and MP3 CD: A Textbook for Beginning Arabic*, Part Three, Georgetown University Press.

Buckwalter Arabic Morphological Analyzer Version 2.0 (Catalog #: LDC2004L02), http://www.ldc.upenn.edu/Catalog/ByYear.jsp#2004

Cavalli-Sforza, V. & Chekayri, A. (2010), Information Technology Support in Teaching Arabic as a Foreign Language. *Colloque Internationale. Le curriculum de la langue arabe-Choix théoriques et méthodes d'application.* Centre de recherche scientifique et technique pour le développement de la langue arabe, 17-18 May, 2010, Algiers.

Corpus oraux : Essai de segmentation automatique

Noura Tigziri
Département de langue et culture amazighes
Université Mouloud Mammeri de Tizi-Ouzou

Nora.tigziri@gmail.com

Introduction : Présentation du projet

Notre projet consiste en la mise en place d'une banque de données de corpus oraux, numérisés, transcrits et annotés pour la langue amazighe qui soit exploitable à des fins scientifiques s'adressant principalement aux enseignants chercheurs linguistes. Nous souhaitons récolter un corpus suffisamment large pour qu'il soit représentatif de la langue, et afin qu'il permette sa sauvegarde sous forme de ressource linguistique. Cette recherche fait intervenir deux institutions : le département de langue et culture de Tizi-Ouzou et la section linguistique de la Faculté de lettres de l'université de Lausanne. Aucun moyen financier spécifique n'accompagne ce projet mais ce dernier a été intégré dans le laboratoire de recherche « Aménagement et enseignement de la langue amazighe » agrée en 2009.

Les objectifs :

Le premier objectif est de mettre à disposition de linguistes une ressource linguistique ce qui implique des conséquences sur la manière de définir les métadonnées et les annotations. Cette recherche est aussi une occasion de documenter le kabyle parlé dans toutes ses variétés, sous toutes ses formes géographiques. Son intérêt réside aussi du fait que cette ressource linguistique sera accessible via le web. Ainsi, on peut ajouter que cette banque de corpus n'a pas pour objectif le TAL ou le TIC mais un outil aussi complet possible (métadonnées, annotations, étiquetage…) pour des linguistes qui pourraient s'intéresser à un ou des élément(s) de recherche.

La création d'un corpus oral, se fait sur la base l'article de Jacobson (2002), chercheur au LACITO (Laboratoire de langues et civilisations à tradition orales). Nous intégrerons l'écrit en utilisant la notation usuelle du kabyle. Les corpus constitués, nous les écrivons en notation usuelle et les retranscrivons en transcription phonétique (API) (Annexe transcription). Cette opération étant faite, nous y ajoutons des métadonnées qui permettront d'identifier nos données et les

décrire (date, langue etc.). Nous nous basons sur les recommandations d'OLAC pour le codage des métadonnées (LACITO,

http://lacito.vjf.cnrs.fr/archivage/index.htm) même si d'autres modèles (ALAVAL,http://www2.unine.ch/dialectologie/page9353.html, CRDO, http://crdo.risc.cnrs.fr/exist/crdo/ et http://crdo.up.univ-aix.fr/) sont aussi intéressants.

La conservation des données se fera grâce à des copies et à la numérisation (transformation en ressource linguistique informatisée). En effet, comme le rappelle Jacobson (2002), le mode de représentation digital a l'avantage d'être répandu, facile d'emploi et a la capacité de mieux conserver les données. Nous utiliserons un codage sans compression pour nos données audio, ce qui semble plus adapté pour l'archivage à long terme.

Le travail sur le terrain :

Pour atteindre notre but nous enregistrons des corpus de locuteurs monolingues. Ces corpus sont recueillis par nos étudiants de licence de notre département. Ceci a un double objectif : - cibler toutes les régions de la Kabylie grâce à eux qui proviennent des quatre coins de notre terrain d'enquête.
– compléter la formation de nos étudiants. Des consignes strictes sont données aux enquêteurs : Faire transcrire le même corpus par deux étudiants, indépendamment l'un de l'autre. Un membre de l'équipe comparera ensuite ces deux transcriptions pour repérer d'éventuelles écarts récurrents (par exemple variation fréquente entre [k] et [t], entre occlusive et spirante etc.) qui peuvent être l'indice de difficultés. Contrôler toutes les transcriptions faites par les étudiants indépendamment par deux
membres de l'équipe (avec réécoute de l'enregistrement simultanément) et la faire évaluer grossièrement (par exemple: Très bon - Bon - Suffisant - Insuffisant). On comparera ensuite les évaluations données et on réexaminera les transcriptions pour lesquelles les évaluations diffèrent de façon importante (de plus d'un degré). On réexaminera également toutes les transcriptions jugées insuffisantes par un évaluateur au moins pour décider de celles qui devraient être écartées comme trop fautives et refaites. On identifiera clairement quels étudiants ont transcrit quels corpus, quels membres de l'équipe l'ont contrôlé et conserver cette information (ce seront des métadonnées importantes). Il pourrait être utile d'avoir des informations de type sociolinguistique sur les étudiants qui transcrivent...

Nous avons établi pour chaque locuteur une fiche de collecte (Annexe 1) où doivent apparaître les métadonnées préalablement définies. Pour compléter ces données, nous avons établi des listes de mots (Annexe 2) en fonction de plusieurs paramètres dont les différents champs sémantiques que nous soumettons dans les divers points d'enquête.

Choix technologiques :

Nous avons opté pour l'adoption de standarts (OLAC ; xml) et des logiciels autant que possible gratuits, open-source et multi-plateformes (Windows-Mac_OSX-Linux).

Le traitement et l'informatisation des corpus oraux supposent un certain nombre d'outils théoriques et de techniques qu'on devait maîtriser. Le premier point est la définition des métadonnées. La question des métadonnées commence à se poser sérieusement lorsque se multiplient les ressources linguistiques informatisées et potentiellement accessibles en ligne. Il s'agit de se mettre d'accord sur des descripteurs qui permettront ensuite une recherche efficace dans un catalogue qui renverra aux ressources elles-mêmes.

Dans la constitution d'un système de métadonnées pour des données - ou "ressources" - linguistiques (enregistrements audio ou video, photos, transcriptions, annotations), différents niveaux peuvent être considérés:

- *Description générale de la ressource linguistique (langue, variété, date de recueil, genre…)*
- *Description des traits spécifiques de la ressource linguistique (date, lieu, enquêteur, informateur, moyens techniques, fichiers (noms, types, localisation…)*

Pour notre projet cela nous concerne

1) puisque l'un des objectifs, dans l'avenir, est de publier les informations sur les ressources construites pour permettre à d'autres chercheurs de savoir qu'elles existent et, le cas échéant, d'y accéder. (Mais rendre publiques les métadonnées n'impliquent pas obligatoirement de rendre l'accès à ces données également libre)

2) Comme il est prévu un grand nombre de corpus élémentaires (=enregistrements ou sessions...) il faut alors, de toutes façons, se construire un système de métadonnées pour retrouver rapidement un sous-ensemble de données. Alors autant le construire de façon à ce qu'il soit compatible avec un système standardisé.

En relation avec ces structures de métadonnées des logiciels capables de les utiliser ont été développés (OLAC, IMDI…)

Pour notre part nous avons choisi d'utiliser OLAC (http://linguistlist.org/olac/index.html)

L'OLAC a élaboré son système de métadonnées pour la description de ressources linguistiques. Il est simple et assez général, mais la formalisation d'un mécanisme

d'extensions permet d'être plus spécifique.

Pour notre recherche l'examen, même rapide, de ces systèmes de métadonnées a eu le mérite de nous permettre de contrôler que rien d'essentiel n'a échappé à notre projet de "fiche de collecte". On voit ainsi, par exemple, que cette fiche ne permet pas de décrire le genre de données recueillies: soliloque, conversation, réponses orales à des questions, poèmes etc...

D'autre part les notations de lieux (d'enquête, de naissance etc) devraient être précisées par une indication longitude/latitude en raison du grand nombre de noms de lieux identiques - donc ambigus - en Kabylie.

Actuellement, nous sommes arrivés à 700 points d'enquête, et 400 enregistrements de 20mn chacun pour la plupart transcrits (Annexe 3 : exemple de corpus). Nous avons établi une « carte exemple » d'un certain nombre de points d'enquête (Annexe 4).

Nous avons, pour le moment utilisé Google Earth pour la représentation spatiale de ces points d'enquête ; La définition des coordonnées de ces points (longitude et latitude) n'a pas été une tâche facile. En effet, les toponymes présentent une grande variation dans le temps et dans l'espace. Il nous arrive de ne pas pouvoir situer exactement un point d'enquête sur la carte parce le nom a changé ou a été transformé. En effet, les diverses sources (cartes topographiques, enquêtes de Basset, documents administratifs fournis par la Wilaya) présentent parfois, des variations importantes dans les toponymes et ceci est une difficulté supplémentaire à surmonter quand on passe à une représentation cartographique.

Enrichissement des données :

La première opération indispensable pour passer de corpus oraux au corpus écrits est la préparation d'un clavier qui pourrait nous faciliter l'utilisation des caractères spécifiques du kabyle. Pour ce faire nous sommes partis des conventions d'écriture de l'INALCO (http://www.inalco.fr/crb/pages_htmel/tableau_prononciation_kab.html) et UNICODE pour élaborer ce clavier.

Unicode c'est fantastique parce qu'on peut utiliser des dizaines de milliers de caractères dans une seule police...

Mais Unicode c'est infernal parce qu'on peut réaliser la même lettre de plusieurs façons différentes et que ces différences, si elles ne sont pas toujours facilement perçues par l'oeil humain, sont un abîme pour un ordinateur.

Le problème se pose pour les caractères complexes (notation des emphatiques par exemple) qui peuvent exister comme caractères uniques en quelque sorte pré-

construits et occupant une position dans la grille Unicode ou bien être produit par l'association de deux caractères: une lettre et un signe diacritique.

Or les programmes informatiques vont traiter différemment ces deux situations. Les logiciels permettant d'établir des listes de fréquence ou des concordances fonctionnent correctement lorsque les caractères complexes sont codés par des caractères uniques mais ne savent pas traiter le cas où ils sont formés par l'association de deux caractères.

La règle à appliquer est donc la suivante: si c'est possible, écrire un caractère complexe en utilisant un caractère unique et non pas en combinant un caractère littéral et un caractère diacritique.

C'est pourtant exactement l'inverse que propose le site edition.berbere...

(http://edition.berbere.free.fr/tables_saisie_berbere_utf-8_01.html).

Les propositions qui figurent dans le tableau suivant respectent la règle ci-dessus

Pour les consonnes labiovélaires (pas reprises ici) il n'y a pas de caractères uniques dans Unicode. La proposition de l'INALCO - postposition de ° - reste donc la plus simple puisque ° en exposant se trouve directement sur tous les claviers.

	Bloc	Code		Bloc	Code
a	Latin de base	0061	A	Latin de base	0041
b	Latin de base	0062	B	Latin de base	0042
c	Latin de base	0063	C	Latin de base	0043
č	**Latin étendu-A**	**010D**	**Č**	**Latin étendu-A**	**010C**
d	Latin de base	0064	D	Latin de base	0044
ḍ	**Latin étendu suppl.**	**1E0D**	**Ḍ**	**Latin étendu suppl.**	**1E0C**
e	Latin de base	0065	E	Latin de base	0045
f	Latin de base	0066	F	Latin de base	0046
g	Latin de base	0067	G	Latin de base	0047
ǧ	**Latin étendu-B**	**01E7**	**Ǧ**	**Latin étendu-B**	**01E6**

(NPC avec ğ et Ğ [diacritique : brève !] de Latin étendu-A 011F et 011E)

h	Latin de base	0068	H	Latin de base	0048
ḥ	**Latin étendu suppl.**	**1E25**	**Ḥ**	**Latin étendu suppl.**	**1E24**
i	Latin de base	0069	I	Latin de base	0049
j	Latin de base	006A	J	Latin de base	004A
k	Latin de base	006B	K	Latin de base	004B
l	Latin de base	006C	L	Latin de base	004C
m	Latin de base	006D	M	Latin de base	004D
n	Latin de base	006E	N	Latin de base	004E
ɣ	**Extensions IPA**	**0263**	**Ɣ**	**Latin étendu-B**	**0194**

(Attention ! autre possibilité : bloc Grec et Copte avec le couple γ 03B3 *pour la minuscule et* Γ 0393 *pour la capitale. On pourrait réserver ces caractères, si nécessaire, à la notation d'une réalisation spirante d'un /g/)*

q	Latin de base	0071	Q	Latin de base	0051
r	Latin de base	0072	R	Latin de base	0052
ṛ	**Latin étendu suppl.**	**1E5B**	**Ṛ**	**Latin étendu suppl.**	**1E5A**
s	Latin de base	0073	S	Latin de base	0053
ṣ	**Latin étendu suppl.**	**1E63**	**Ṣ**	**Latin étendu suppl.**	**1E62**
t	Latin de base	0074	T	Latin de base	0054
ṭ	**Latin étendu suppl.**	**1E6D**	**Ṭ**	**Latin étendu suppl.**	**1E6C**
ţ	**Latin étendu-A**	**0163**	**Ţ**	**Latin étendu-A**	**0162**
u	Latin de base	0075	U	Latin de base	0055

w	Latin de base	0077	W	Latin de base	0057
x	Latin de base	0078	X	Latin de base	0058
y	Latin de base	0079	Y	Latin de base	0059
z	Latin de base	007A	Z	Latin de base	005A
ẓ	**Latin étendu suppl.**	**1E93**	Ẓ	**Latin étendu suppl.**	**1E92**
ɛ	**Extensions IPA**	**025B**	Ɛ	**Latin étendu-B**	**0190**

(Attention ! d'autres possibilités seraient envisageables, p.ex. bloc Grec et Copte...)

Généralisation à la notation de la spirantisation.

Le principe consistant à préférer systématiquement l'utilisation d'un caractère unique sur l'association de deux caractères est également préférable pour les autres niveaux de transcription. Ainsi, pour la notation des spirantes, dans une transcription phonétique large, si l'on décide d'adopter la convention du trait souscrit (suscrit sur g ou G) plutôt que le recours aux caractères de l'API, il vaudra mieux utiliser les caractères qui apparaissent dans le bloc Latin étendu supplémentaire plutôt que de combiner un caractère avec le diacritique « trait souscrit » (Unicode 0320).

Pour écrire le kabyle, en plus des lettres habituelles on a besoin:

- des lettres: ɣ et ɛ
- des lettres diacritées: c et g avec caron (appelé encore: hacek, chevron, antiflexe, accent hirondelle, v suscrit), d, h, r, s, t, z avec point souscrit, t cédille.

Tous ces caractères doivent être disponibles en lettres minuscules et en lettres capitales (majuscules).

Toutes ces lettres sont prévues, précomposées, dans divers blocs Unicode. Les codes correspondants sont indiqués ci-dessus.

Pour permettre la saisie de ces lettres sans exiger de trop gros efforts de mémorisation et éviter des conflits avec des combinaisons de touches prédéfinies par le système ou par d'autres programmes (Word par exemple), la solution

générale retenue consiste à définir une "touche morte", au fonctionnement analogue à la touche de l'accent circonflexe ou du tréma. On presse la touche morte puis la touche correspondant portant un caractère simple (dit "de base") et on obtient le caractère spécial voulu.

La touche retenue comme touche morte est celle qui, sur le clavier suisse romand, porte les signes < et >.

Le ɣ et le ɛ s'obtiennent avec la touche morte suivie des touches y et e (les caractères de base les plus proches par leur forme).

Les ḍ , ḥ , ṛ , ṣ , ṭ et ẓ avec la touche morte suivie des caractères de base correspondants d, h, r, s, t et z.

Pour ţ la touche morte est suivie de la touche x ("i*ks*").

Les lettres capitales s'obtiennent normalement en combinant la touche morte avec la touche shift (majuscule).

Les caractères < et > restent disponibles: il suffit de les taper après la touche morte: deux pressions successives sur la touche < donne < ou > si la touche shift est pressée.

Concrètement, et suivant Sur Macintosh (avec clavier Français-Suisse), ou sur PC, il faut suivre les opérations suivantes :

Sur Macintosh (avec clavier Français-Suisse)

- installer le fichier +kabyle.keylayout (créé avec le logiciel gratuit Ukelele cf. scripts.sil.org/ukelele) dans le dossier Keyboard Layouts qui se trouve dans le dossier Bibliothèque (ou: Library) de l'utilisateur (ou de l'ordinateur iMac). (Si le dossier Keyboard Layouts n'existe pas il faut le créer dans le dossier bibliothèque, en lui donnant exactement ce nom);

- redémarrer l'ordinateur;

- ouvrir les Préférences Système... (menu Pomme) et ensuite International; cliquer sur l'onglet Menu Saisie, rechercher le clavier +kabyle et cocher la case à gauche (Activé);

- dans la barre en haut de la fenêtre, à droite, cliquer sur le drapeau qui symbolise le clavier (combinaison des drapeaux suisse et français) et sélectionner le clavier +kabyle qui doit se trouver en dessous.

Le clavier +kabyle est désormais accessible et toute application utilisant une police

Unicode assez complète (comme Doulos SIL) permettra d'obtenir les caractères spécifiques nécessaires avec la touche morte.

Mais il y a un logiciel gratuit, Microsoft Keboard Layout Creator, accessible ici: http://www.microsoft.com/globaldev/tools/msklc.mspx, qui permet de reconfigurer un clavier et, notamment, de créer une touche morte.

Toutefois l'utilisation de ce programme exige l'installation préalable de l'environnement de programmation .NET (.NET Framework) à télécharger ici :

http://www.microsoft.com/net/Download.aspx

Un mode d'emploi en français ci-joint (MKLC_fr.pdf; extrait de http://llacan.vjf.cnrs.fr/fichiers/manuels/Internet/SaisieClavier.pdf) permet de se débrouiller assez facilement. Il faut simplement corriger ce qui est dit sur l'installation du clavier dans Windows:

- le fichier .msi est le fichier composé du nom du clavier et de l'abréviation de la famille du processeur (le plus souvent i386). Mais il y a un fichier de Setup qui doit se charger d'installer la bonne version. Attendre le message: Installation complète. Ce n'est pas immédiat.

- c'est le panneau de configuration Options régionales (et non Clavier) qui, sous Windows XP en tout cas, permet d'installer et d'activer le nouveau clavier.

Bien entendu il est possible de choisir n'importe quelle touche comme touche morte, pas seulement le <.

Extension envisageable:

Si on le souhaite, on peut ajouter d'autres caractères, comme par exemple le ə (<+a). La difficulté consiste à attribuer les caractères supplémentaires à une touche présentant, si possible, un certain rapport, pour éviter un effort de mémoire. Mais on pourrait parfaitement, par exemple, définir une autre touche morte pour entrer des caractères de l'alphabet phonétique utilisés dans une transcription phonético-phonologique. On pourrait avoir, par exemple, avec $ comme touche morte $+t donnant θ, $+d donnant δ etc.

Sites cités:

- pour télécharger Ukelele, logiciel de configuration du clavier pour Macintosh:

http://scripts.sil.org/ukelele

- pour télécharger Microsoft Keboard Layout Creator, logiciel de configuration du

clavier pour PC:

http://www.microsoft.com/globaldev/tools/msklc.mspx

- pour télécharger l'environnement de programmation .NET pour Windows, s'il n'est pas installé: http://www.microsoft.com/net/Download.aspx

Enrichissement des données :

L'enrichissement des données par un certain nombre de logiciels présuppose la mise en place d'un certain nombre de concepts qui pourraient nous aider dans la segmentation des corpus en unités (énoncés, phrase…) et l'étiquetage linguistique (morphosyntaxique). L'un des points qui nous intéressent est la relation phrase/prosodie/segmentation.

Philippe Martin (1981, 2002, 2010) définit assez clairement les concepts qui nous intéressent pour notre problématique. Ainsi, pour lui, le mot prosodique est l'unité prosodique minimale contenant un seul mot accentué. Cela correspond, généralement, au syntagme. Ceci explique la composition du groupe prosodique de mots prosodiques. Quant à la phrase prosodique, toujours d'après Philippe Martin, elle indique la courbe mélodique phrastique, dépendant de la modalité de la phrase (déclarative, interrogative etc.).

Dans cette perspective, la phrase prosodique n'est qu'une suite de mots délimitée par deux pauses importantes (initiale et finale) et caractérisée par une intonation qui varie avec le type de phrase (assertive, interrogative, injonctive).

Pour la définition de la phrase et de l'énoncé, le Dictionnaire de la linguistique de Georges Mounin (2004 : 262) stipule "Beaucoup d'usages linguistiques tiennent *énoncé* et *phrase* pour des termes synonymes. Mais on a intérêt à opposer les phrases (unités de langue) aux énoncés (unités ou exemples de parole), l'énoncé étant ce qui est donné dans le matériau non analysé".

Un énoncé est "tout segment de la chaîne parlée, compris entre deux interruptions nées soit du silence, soit du changement de locuteur, et qui n'a pas encore été identifié ou analysé en phrases" (G. Mounin, 2004 : 125).

La phrase est définie par A.Martin (1991 : 131) comme une séquence *«dont tous les éléments se rattachent à un prédicat unique ou à plusieurs prédicats coordonnés »*. Pour la syntaxe, il déclare (1985 :13) «S'il est un point sur lequel peuvent tomber d'accord les linguistes contemporains, à quelque école qu'ils se rattachent, c'est qu'appartient à la syntaxe l'examen de la façon dont les unités linguistiques douées de sens se combinent, dans la chaîne parlée, pour former des énoncés (…) c'est-à-dire la façon d'ordonner des mots pour former des phrases». Et l'objet de la syntaxe est «d'exprimer par quels moyens les rapports qui existent entre les éléments d'une expérience(…) peuvent être marqués dans une succession

d'unités linguistiques de manière que le récepteur du message puisse reconstruire cette expérience» (Ibid., 2-8, p.16)

Toujours pour notre étiquetage linguistique, nous avons un élément aussi important que la phrase et l'énoncé, à savoir le syntagme pour qui la définition de A.Martinet est tout à fait indiquée puisqu'il le définit (Martinet, 1980, 4-13, p.112) comme « toute combinaison de monèmes dont les rapports mutuels sont plus étroits que ceux qu'ils entretiennent avec les autres éléments de l'énoncé, plus, éventuellement, le monème fonctionnel qui rattache cette combinaison au reste de l'énoncé»

Dans un énoncé complexe, on trouve donc un énoncé minimum qui se compose généralement d'un prédicat et d'un sujet (expansion obligatoire) et les expansions. Le prédicat est l'élément irréductible de l'énoncé.

Donc, une phrase est un énoncé dont tous les éléments se rattachent à un prédicat unique ou à plusieurs prédicats coordonnés en tenant compte des pauses importantes (initiale et finale) de l'intonation qui varie avec le type de phrase (assertive, interrogative, injonctive).

Deux logiciels gratuits et libres d'accès sont utilisés : PRAAT(http://www.fon.hum.uva.nl/praat/download_win.html)

et JAXE (http://sourceforge.net/projects/jaxe/)

JAXE :

Un langage XML est défini de façon formelle, de manière à permettre la vérification automatique de la syntaxe. Cette description formelle comprend les noms des éléments du langage, les imbrications possibles entre les éléments, l'ordre autorisé des éléments, et leurs attributs (les attributs étant optionnels ou obligatoires).

Jaxe facilite la création des documents XML en utilisant les règles du langage pour proposer des éléments à insérer là où c'est possible. Cela permet de créer des documents *valides* (c'est à dire se conformant aux règles du langage) beaucoup plus facilement qu'avec un simple éditeur de texte.

En plus des fichiers décrivant les langages XML (les `schémas`), Jaxe utilise des *fichiers de configuration* qui définissent la barre de menus et la façon d'afficher les éléments du langage. Ces fichiers se trouvent dans le répertoire `config`, et leur nom se termine en `_config.xml`.

La composante « Métadonnées »

Chaque enregistrement est accompagné d'une « fiche de collecte » qui le décrit. Cette fiche de collecte :

- Sera intégralement transcrite sous la forme d'un fichier xml (éléments : Divers, Enquêté, Collecteur, Debriefing, Autres infos ; sous-éléments : les différentes lignes de la fiche), auquel il sera référence dans le document élémentaire ;

- Constituera la source des métadonnées incluses dans le document élémentaire.

  Le choix des métadonnées retenues pour accompagner directement chaque document élémentaire se fonde sur les standarts reçus (Dublin Core (DC) et OpenLangage Archives Community (OLAC) et suit d'assez près les recommandations du Centre de Ressources pour la Description de l'Oral (CRDO, CNRS). On renonce cependant à noter les caractéristiques constantes de nos documents : la langue étudiée (le kabyle) et la langue d'étude (le français)

- On donne, sous-l'élément Métadonnées, la liste des sous-éléments (= représentation de la structure hiérarchique)

- On définit ensuite, comme des éléments distincts, en dehors de la spécification de l'élément Métadonnées, chacun de ces sous-éléments (= représentation des composantes de la structure). Ces sous-éléments de l'élément Métadonnées sont donc des éléments et peuvent à leur tour se composer de sous-éléments.

Cette représentation est donnée en Annexe 5

La composante de l'élément Données

La première composante est la Phrase qui va être analysée en une succession de parties du discours et qui peut-être glosée (traduction juxtallinéaire), transcrite, en phonétique ou en phonologie, traduite. Elle est également liée à un élément sonore.
Les éléments de glose, de transcriptions, de traduction et de lien avec le signal audio caractérisent également les différentes parties du discours. C'est pourquoi on les réunit en un « ensemble » (nommé, dans cet exemple, « formes »).
Les parties du discours, ainsi que les attributs qui les caractérisent, sont déterminées par les linguistes berbérisants du groupe de recherche.
Pour que le fichier de description de la structure soit accepté par Jaxe, il faut encore indiquer un élément racine de l'arborescence hiérarchique. Dans notre exemple ce sera l'élément Document_kabyle.
L'illustration est en Annexe 6, 7, 8, 9.

Application avec PRAAT :

PRAAT () est exploité en analyse acoustique. En créant de nombreuses tires, on arrive aligner le signal temporel, le sonagramme, la notation usuelle, le découpage en unités préalablement définies ou étiquetage linguistique (racines, schèmes, syntagmes…) (Annexe 10). Des scripts sont également utilisés à des fins de segmentation en énoncés par exemple. Evidemment toute la problématique de la

définition de l'énoncé en ce qui concerne l'oral est difficilement maîtrisable. Pour notre part, les pauses sont prises comme indicateur de séparations d'énoncés (Annexe 11, 12). Evidemment PRAAT a aussi la qualité d'aligner son/transcription.

Annexe 1 :

Fiche de collecte

1. divers		
date de collecte :	2009	
lieu :	Tigzirt (Tasalast et Tamda Ouguemoune)	
support de l'enregistrement :		
durée de l'enregistrement :	Environ 45 minutes	
lieu de l'enregistrement :	Au bord de la mer	
sujet de l'enregistrement :	Poissons, animaux de la mer,	
Y avait-il un public ?	Non	
Référence		
2. enquêté		
(Nom :)		
Date de naissance :	L'un est né en 1934 / l'autre est né en 1977	
Sexe :	Hommes	
Village d'origine :	Tigzirt	
Tribu :	Iflissen	
Domicile actuel (village, région):	Tigzirt	
Dialecte parlé, (nom donné par le locuteur à son parler)	Kabyle	
Autre (s) langue (s) parlée (s) :	Kabyle, arabe	
(Au travail :)	?	

(À la maison :)	Kabyle	
Séjour (s) à l'étranger	Non	
Durée du/des séjour(s)	?	
Scolarité et formation	Niveau CEM celui qui est né en 1977	
Langue(s) de l'enseignement reçu :		
Profession :	Chasseurs marins	
Personne(s) ayant joué un rôle dans l'apprentissage linguistique (par exemple son père, sa mère, personne avec qui le locuteur a passé son enfance)		
- lien de parenté, relation avec la personne :	Non	
- lieu d'origine :		
- scolarité (et langues d'enseignement) :		
situation familiale (mariage(s), enfants) :	Marié (celui qui est né en 1934), célibataire (celui qui est né en 1977)	
langue (s) parlée (s) par le conjoint :	Kabyle	
attitude du locuteur par rapport à sa langue et à sa façon de parler :	Fière vis-à-vis du kabyle, leur parler différent au reste de la Kabylie ;	
3. Collecteur		
nom, prénom:	Oumaouche Omar	

langue (s) parlée (s) :	Kabyle, arabe, français	
origine :	Tigzirt	
relation enquêteur-enquêté :	?	
4. Debriefing		
conscience du micro :		
attitude du locuteur par rapport à l'enregistrement :		
attitude du locuteur par rapport à l'entretien, aux questions posées…		
5. Autres infos		

Annexe 2 :

mot (en français)	**Parler 01 : Aglala / Beni Zmenzer**	**Parler 02 : Isseradjène / Boudjima**
Champignon	*Tireγlin Racine : RΓ L Schème: $tic_1ec_2c_3ic_4$	*Tireγla Racine : RΓ L Schème : $tic_1ec_2c_3a$
Petits pois	*Tajijbant R : JBN S : $tac_1ic_2c_3ac_4t$	*Tajilbant R : JLBN S : $tac_1ic_2c_3ac_4t$
Pin Silvestre	*Tazumbit R : ZMB S : $tac_1uc_2c_3it$	*Tazumbilt R : ZMBL S : $tac_1uc_2c_3ic_4t$
Citronnier	*Tilimet R : LM S : tic_1ic_2et	*Talimet R : LM S : tac_1ic_2et
Lentisque	*Imidek R : (m) DK S : $ic_1ic_2ec_3$	*Tidekt R : DK S : $ti\ c_1ec_2t$
Lentilles	*Laɛ des R: □DS S : $c_1ac_2c_3ec_4$	*Lɛ eḍ s R : □Ḍ S S : $c_1c_2ec_3c_4$
Eucalyptus	*Karitus R : KRTS S : $c_1ac_2ic_3u\ c_4$	*Akalatus R : KLTS S : $ac_1ac_2ac_3uc_4$
Chêne liège	*Akerruc R : KRC S :a $c_1eC_2uc_3$	*Ikirruc R : KRC S : $ic_1iC_2uc_3$
Gland	****Aḥ elluḍ*** ***R :*** Ḥ ***L***Ḍ ***S : $ac_1eC_2uc_3$***	****Abelluḍ*** ***R : BL***Ḍ ***S : $ac_1eC_2uc_3$***
Rue	*Awermi R : WRM S : $ac_1ec_2c_3i$	*Lfenğla R : LFĞL S : $c_1c_2ec_3c_4c_5a$

Rosier	*Tiɛ fart : R : Γ FR S : $tic_1c_2ac_3t$	*Taɛ fart R : Γ FR S : $tac_1c_2ac_3t$
Palmier	*Tazdayt R : ZDY S : $tac_1c_2a\ c_3t$	*Tazanet R : ZN S : tac_1ac_2et
Citrouille	*Taxsayt R : XSY S : $tac_1c_2a\ c_3t$	*Taxsakt R : XSK S : $tac_1c_2ac_3t$
Luzerne	*Ikeffis R : KFS S : $ic_1eC_2ic_3$	*Ikeffil R : KFL S : $ic_1eC_2ic_3$
Figuier de barbarie	*Lkermus R :(L) KRMS S :$c_1c_2e\ c_3c_4uc_5$	*Lkermus R: (L) KRMS S :$c_1c_2ec_3c_4uc_5$
Mûres sauvages	* Tinijwal R : NJWL S : $tic_1ic_2c_3ac_4$	*Tizwal R: ZWL S : $tic_1c_2ac_3$
	* Timendekrar R : MNDKR S : $tic_1ec_2c_3ec_4c_5ac_6$	*Tiferkekkay R : FRKY S : $tic_1ec_2c_3eC_4ac_5$
	*Timeccucin R : MC S :$tic_1eC_2uc_3ic_4$	*Aɛersiwen R : Γ RṢ W S : $ac_1ec_2c_3ic_4ec_5$
Lait	*Ayefki R : YFK S : $ac_1ec_2c_3i$	*Ifki R : FK S : $i\ c_1c_2i$
Frère	*Xuya R : XY S : c_1uc_2a	*Ugma R : GM S : uc_1c_2a
L'argent donné pour la fiancée	*Tiẓri R: ẒR S : $ti\ c_1c_2i$	*Tiẓri R : ẒR S : tic_1c_2i
Filles	*Tullas R : LS S : tuC_1ac_2	*Tiḥdayin R :ḤDY S :$tic_1c_2ac_3ic_4$
Amas de bois qui sert à cuire la poterie	*Uγud R : ΓD	*Uγud R : ΓD

	S : u c_1uc_2	S : uc_1uc_2
Cruche	*Acmux R : CMX S : $ac_1c_2uc_3$	*Asagem R : GM S : $ac_1ac_2ec_3$
Chapelet de morceau de viande	*Iceddiwen R: CDW S : i $c_1C_2ic_3ec_4$	*lmeck R : (L) MCK S : $c_1c_2ec_3c_4$
Un spécialiste pour circoncire les garçons	*aḥeğğ***am*** ***R :*** ḤĞ***M*** ***S : $ac_1eC_2ac_3$***	*lemɛellem R : (L) ƐLM S : $c_1ec_2c_3eC_4ec_5$
Les œufs à la semoule	*Timcewwect R : MCW S : $tic_1c_2eC_3ec_4t$	*Tabeɣrirt R : BΓR S : $tac_1ec_2c_3ic_4t$
Bébé	*Llufan R : LFN S : $C_1uc_2ac_3$	*Agrud R : GRD S : $ac_1c_2uc_3$
Crêpes	*Lemsemmen R : (L) MSMS S :$c_1ec_2c_3eC_4ec_5$	*Aḥ***eddur*** ***R :***Ḥ***DR*** ***S : a $c_1eC_2uc_3$***
Petite fille	*Taqcict R : QC S :$tac_1c_2ic_3t$	*Tagruḍt R : GRḌ S : $tac_1c_2uc_3$
Faire purifier, circoncire clarifier	*Sḍehren R :ḌHR S :$c_1c_2ec_3c_4ec_5$	*Zeyynen R : ZYN S :$c_1e\,C_2c_3ec_4$
Faire manger	*Aceččі R : C S : ac_1eC_2i	*Aseččі R : C S : ac_1eC_2i
Enfant	*Aqcic R : QC S : $ac_1c_2i\,c_3$	*Aqcic R : QC S : $ac_1c_2ic_3$
A ce moment là	*Imir-n R : MR S : $ic_1ic_2c_3$	*Imir R : MR S : ic_1ic_2
Souhait	*Saɛd R : SƐd	*Henni R : HN

	S : $c_1a\,c_2c_3$	S : c_1eC_2i
Je lui dois, elle leur doit	*Tettalas R : LS S : $teC_1ac_2ac_3$	*Tettaras R : RS S : $teC_1ac_2ac_3$
Peuvent	*Zemren R ZMR S : $c_1ec_2c_3ec_4$	*Waεan R : Wε S : $c_1ac_2ac_3$
Haut de robe qui bouffe en poche au dessus de la ceinture	*Icimmi R : CM S : $ic_1i\,C_2i$	*Iciwi R : CW S : ic_1ic_2i
La cuise	*Taγma R : ΓM S : tac_1c_2a	*Taqesbuḍt R: QSBḌ S : $tac_1ec_2c_3uc_4$
Celles qui roule la semoule avec les mains dans un grand plat pour la préparation du couscous	*Tifettalin R : FTL S : $tic_1eC_2a\,c_3ic_4$	*Tineffalin R : NFL S : $tic_1e\,C_2a\,c_3ic_4$
Etre d'accord	*Mseqbalen R :(MS) QBL S : $c_1c_2ec_3c_4ac_5ec_6$	***mruḍan** R: (M)RḌ S :$c_1c_2uc_3ac_4$
Marie	*Isli R : SL S : ic_1c_2i	*Isli R : SL S : ic_1c_2i
Ce qu'il faut	*Ilaqen R : LQ S : $ic_1ac_2ec_3$	*Ilezmen R : LZM S : $ic_1ec_2c_3ec_4$
Faire les youyous	*Siγret R : ΓRT S : $c_1ic_2c_3ec_4$	*Seγret R : ΓRT S : $c_1ec_2c_3ec_4$
Beignets	*Lesfenğ R : (L) SFNĞ S : $c_1ec_2c_3ec_4c_5$	*Lexfaf R (L) XF S : $c_1ec_2c_3ac_4$
Insectes	*Ibaεεac R : BεC S : $ic_1aC_2ac_3$	*Ibelεac R : BLε S : $ic_1ec_2c_3ac_4$
	*Aqrur	*Agrud

Enfant	R : QR S :$ac_1c_2uc_3$	R : GRD S : $ac_1c_2uc_3$
Les enfants	*Arrac R : RC S : aC_1ac_2	*Igerdan R : GRD S : $ic_1ec_2c_3ac_4$
Ensemble	*Ilkelli R : LKL S : $ic_1c_2eC_3i$	*Urkelli R : RKL S :$uc_1c_2eC_3i$
Faire partie du cortège qui chercher la marie	*Iεerrasen R :ƐRS S : $ic_1eC_2ac_3ec_4$	*Iqeffafen R : QF S : $ic_1eC_2ac_3ec_4$
Toute petite	*Taṭuṭaḥt R :ṬḤ S : $tac_1uc_2ac_3t$	*Tamecṭuḥt R :MCṬḤ S : $tac_1ec_2c_3uc_4t$
Bébé	*Ṣṣebyan R :ṢBY S : $C_1ec_2c_3ac_4$	*Agrud R : GRD S : $ac_1c_2uc_3$
Veau	*Aεejmi R : ƐJM S : $ac_1ec_2c_3i$	*Agenduz R : GNDZ S : $ac_1ec_2c_3uc_4$
Achille gold	*Γlilu R :ΓL S :$c_1c_2ic_3u$	*Qlilu R : QL S : $c_1c_2ic_3u$

Cigale	*Zḍeğ R : ZḌĞ S :$c_1c_2ec_3$	*Tejḍeč R : JḌČ S : $tec_1c_2ec_3$
Chouette	*Timiεruft R : MƐRF S : $tic_1i\ c_2c_3uc_4t$	*Imiεruf R : MƐRF S : $ic_1i\ c_2c_3uc_4$
Hirondelle	*Tifilellest R : FLS S : $tic_1ic_2eC_3ec_4t$	*Tifirellest R : FRLS S : $tic_1ic_2eC_3ec_4t$
Papillon du jour	*Timecriwect R : MCRWC S : $tic_1ec_2c_3ic_4ec_5t$	*Aferṭiṭu R : FRṬ S : $ac_1ec_2c_3ic_4u$

Papillon de nuit	*Aferṭeṭu R : FRṬ S : $ac_1ec_2c_3ec_4u$	*Aferṭiṭu R : FRṬ S : $ac_1ec_2c_3ic_4u$
Singe	*Iddew R : DW S : iC_1ec_2	*Ibekki R : BK S : ic_1eC_2i
Chauve-souris	*Ṭirellil R : ṬRL S : $c_1ic_2eC_3ic_4$	*Itirelli R : TRL S : $ic_1ic_2eC_3i$
Scorpion	*Tiγirdemt R : ΓRDM S : $tic_1ic_2c_3ec_4t$	*Tawejjiḍt R : WJḌ ***S : $tac_1eC_2ic_3$***
Tordeuse	*Aburebbu R : RB S :$abuc_1C_2u$	*burebbu R :RB S :buc_1C_2u

Taon	*Taggent R : GN S : taC_1ec_2t	*Aggen R : GN S : aC_1ec_2
Faucon	*Afalku R : FLK S : $ac_1ac_2c_3u$	*Lbaz R : LBZ S : $c_1c_2ac_3$
Bon plat	*Taḥluqt R : ḤLQ S : $tac_1c_2uc_3t$	*Tarzeft R : RZF S : $tac_1c_2ec_3t$
Jeune pousse	*Issegmi R : SGM S : $iC_1ec_2c_3i$	*Aguḍem R : GḌM S : $ac_1uc_2ec_3$
Renard	*Izirdi R : ZRD S : $ic_1ic_2c_3i$	*Akεab R : KƐB S : $ac_1c_2ac_3$
Vache	*Tafunast R : FNS S : $tac_1uc_2ac_3t$	*Tuwmat R : WM S : tuc_1c_2at
	*Taselluft	*Taselluft

Puce	R : SLF S : $tac_1eC_2uc_3t$	R : SLF S : $tac_1eC_2uc_3t$
Chevale	*Aεewdiw R : ƐWDW S : $ac_1ec_2c_3ic_4$	*Aεidiw R : ƐDW S : $ac_1ic_2ic_3$
Massette (Roseau)	*Agellu R : GL S : ac_1eC_2u	*Tabuda R : BD S : tac_1uc_2a
Figuier	*Tameγrust R : MΓRS S : $tac_1ec_2c_3uc_4t$	*Tanqult R : NQL S : $tac_1c_2uc_3t$
Coquelicot	*Taciḥbuḍt R : CḤBḌ S : $tac_1ic_2c_3uc_4$	*Waḥrir R : WḤR S :$c_1ac_2c_3ic_4$
Plante dont les fruits collent a tous ce qu'elles touchent	*Timenṭeḍt R : MNṬḌ S : $tic_1ec_2c_3ec_4$	*Ihinṭeḍ R : ḤNṬḌ S : $ic_1ic_2c_3ec_4$
Variété de la figue	*Tajenǧalt R : JNǦL S : $tac_1ec_2c_3ac_4t$	*Tajenjirt R : JNJR S : $tac_1ec_2c_3ic_4t$
La vigne	*Ajgagal R : JGL S : $ac_1c_2ac_3ac_4$	*Tajnant R : JN S : $tac_1c_2ac_3t$
Nigelle	*Ssanuǧ R : SNǦ S :$C_1ac_2uc_3$	*Zrareε R :ZR□ S :$c_1c_2ac_3ec_4$
Ver	*Tawekka R : WK S : tac_1eC_2a	*Takečča R : KČ S : tac_1eC_2a
Mouton	*Axerfi R : XRF S : $ac_1ec_2c_3i$	*Ikerri R : KR S :ic_1eC_2i
Escargot	*Aεrus R : ƐRS S : $ac_1c_2uc_3$	*Aεarus R : ƐRS S : $ac_1ac_2uc_3$
Tortu	*Afekrur R : FKR S : $ac_1ec_2c_3uc_4$	*Ifekker R : FKR S :$ic_1e C_2ec_3$

Guêpe	*Areẓ R : RẒ S : ac_1ec_2	*Arẓaẓ R : RẒ S : $ac_1c_2ac_3$
Chevreau	*Aḥuli R :ḤL S : ac_1uc_2i	*Iγid R : ΓD S : ic_1ic_2
Variété de figue	*Abakur R : BKR S : $ac_1ac_2uc_3$	*Abukar R : BKR S : $ac_1uc_2ac_3$
Fenouil	*Lbesbas R: (L) BS S : $c_1c_2ec_3c_4ac_5$	*Abesbas R : BS S : $ac_1ec_2c_3ac_4$
Olivier sauvages	*Aḥeccaḍ R :ḤCḌ S : $ac_1e C_2ac_3$	*Aẓebbuj R : ẒBJ S : $ac_1eC_2uc_3$
Alfa	*Ḥlafa R : ḤLF S : $c_1c_2ac_2a$	*Ḥlafa R : ḤLF S : $c_1c_2ac_3a$

Annexe 3 :

Corpus : Sahel / Bouzeguène

(Transcription en jaxe)

<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>

- <CORPUS>

<CORPUS> <NOTATION USUELLE> : Nekkni zik, ad d-neker deg yiḍ, nṣaf n yiḍ ad nruḥ ta ad teγγar i ta, ad d-nagem d talla n wadda mi i d-newwed ad necγel seksu, mi nfuk seksu-nni, ad nεedi ad nneneḍ leγbar mi nfuk leγbar-nni ad nεedi ad nerfed iqeṭṭaren tasebḥif; ad nruḥ γer cγel. Nettewqam amarḍil Ad naweḍ a yelli γer uzemmur, ad nawi iqcer n uγrum deg yiciwan n naγ; ur nettawi ara llεali-agi i ttawin akka medden tura, wellah ar d tidet a yelli. Ad nawi iqceran-nni n uγrum deg yiciwan nnteγ ad naweḍ aken nemwellah d tislatin d lxalat, deg mi ara nali γef lǧedra alama n fuk-itt-id deg yixef, mi ara d-nars, aεeqqa, ad awḍeγ ar lǧedra ad xezreγ tazemurt ma ufiγ aεeqqa ar teqacuct ad qleγ, ad t-id-yeqḍeγ, ḥemlaγ arrezq a yelli, maci am tura; lǧil n tura. Ad d-nars a yelli tameddit n wass ad aγ-ineγ lazz ad nettḍeqir iqceran-nni n uγrum. "Tecfad yema-m ad tt-ig Rebbi n ṛṛeḥma ; Setti-m ad tt-ig Rebbi n lǧennet. Ula d yema-m tleḥqed". Ad nettḍegir aγrum-nni akka ar sdat ad nluqeḍ, ad nettḍegir aγrum-nni ar sdat ma nufa-d aεeqqa aquran ad t-nsexḍel s uγrum-nni. Ur nesεi reǧwaz ur nesεi, d aγrum-nni kan, ad d-naweḍ s lfarḥ d ameqran ad d-naf tabbarbuct am iqeccaḍen ad tt-neččč d tazidant, d tuzyint. Ad nruḥ ma i yefuk uzemmur-nni d tuga, ma tefuk tuga-nni d nqec n tebḥirin, d timegriwin, d inurrar, d tiεelafin nyezgaren. Ssarwaten madden, deg yiwǧǧiben d tayarza, lxalat d azemmur; irgazen d tiγarziwin n zik d ifellaḥen merra, iεeqliyen ad mexartayen nneεma akk d yirden ;merra ad mxartayen. Kul lexir yettzid imir, kul lexir yettzid. Tura d nkkeẓ i nekkẓen lerẓaq ad imnaε Rebbi lmumen;d nkkeẓ, d nkkeẓ i nekkẓen larẓaq tura. Ur sineγ ara ad hedreγ a tifarγi i yεedan fell-aneγ. Alah, alah nniγ-as a sidi Rebbi ur iγelli yiṭṭij ar d-nfak, wa ar d-nfak, wa sidi, aRebbi ur iγeli yiṭṭij-inna ar d-nfak tirni-inna. Nettfaras; ad teẓred diγ a yewqam-iw asmi wteγ tagut s aḍu, yemut urgaz-iw deg xemsa wetlatin, d amecṭuḥ; yeǧay-d rebεa igerdan, tamurt tella nḥend-ik a Rebbi n cekr- ikTenza tbarquqet, tenza tremant, yenza ifelfel, tenza, deg mi ara d-nekker, tamurt n lefni n ddunit newweḍ-itt. Newweḍ Zubga, at aεbella, newweḍ a yelli Ifarhunen, newweḍ kulci s laḥmul n ukarmus, deg mi ara d-nekkar; hur, hur, hur, ass kamel; hur, hur, hur, ass kamel d tikliwin

ad d-nawweḍ, nezenz, ad d-neččar laḥmul-nni d ker; d lebsel, d lebaṭaṭa, ad nuḥ neznuzuy s yidrimen, tamurt tella a yelli nexdem nečča nḥend-ik a Rebbi n cekr-ik tura wellah a Rebbi ar siniy di leɛmer feyay ar lexla ar akka id ufiy iman-iw ttyimiy akka deg uxxam. Ad nenyec, ad nezdem, ad d-nawi aman di lefni n ddunit mi ara čačarent telliwa ar wasif ara nruḥ deg yiḍ, wellah ar daxel n temdiwin id nettaččar tibettiyin, nettruḥu-d s axxam. I iyeɛedan ihi n leqwanen zik. A zik tabarbuct ma teččiḍ-tt aḥeq Rebbi ar xir n miya u miyin d aksum n tura; tabarbuct nzik akka-tt, timɣarin ma ulac ; ma tfuk teyenat; tekarfat n unebdu ad d-tili lbecna, ma tefuk lbecnatabarkant ad d-tili temelalt, ad d-yili ubeluḍ, d azidan lqut, kulec d azidan d awenɛan. Tura, timɣarin arssant alqec, telmezyin tteddunt ɛaryan a llah ibarek d aya i yellan tura, hata win i yellan. D tidet neznuzuy lleft; yettemɣay-d lleft, nettawi ttɛebga n lleft, imir yelluẓ lḥal, ctaqen medden lqut. Ad neččarttɛebgat n lleft art mura ad ad ɣ-d-fken ablluḍ; d tifrac n ubeluḍ ad t-id-nẓed s tesirt ad d-neggar aɣrum; tabarbuct d taẓidant, aɣrum-is d aẓidan, ticki ulac tifrect-nni ntteks-d azegzaw. Tura, azegzaw tura ma tegreḍ-t-id d aɣrum wellah ma tmenaḍ ad d-yekes deg yimi-k, nettedez amaqcur ad t-negar d aɣrum. A Rebbi di tmexluqt-agi taṛwa taḍsa, tariḍ-ay d iɛegunen aɛɛi, a yamzuy-im. Ikem kan i umi id-hedrey akka wanag lami ad d-hedray ma ur ssiney lehdur. Aḥeq Rebbi ma sney a yelli lehdur, ḥaca ayen yelan akka sufela. Tenna-as Seɛdiya tarihant -ad tt-idker Rebbi s lxir, tmeṭut n Lewnis At cilatt- aken ara ad d-awḍen-ken tinni-as : « ayu;ulac lwexda fell-i alama walay-ten beran-d i yiserxuḍen nsen sddaw uxxam n Juhra n tɛeẓugt ». Di lḥara-nni n Juhra n tɛezugt wina ufella, «ayu, a ysetma timaɛzuzin, ulac lwexda fell-i ad wayiy bran-d i yisarxuḍen nsen seddaw uxxam n Juhra n tɛezugt ». Ad d-awḍen a yelli, lxallat ad ttgejgijent deg yixxamen, ad tent-zuɣu ; zuɣurn-ay amzun d lmal, win weɛan kan ad tsefden s tmegḥelt, win weɛan ad tsefden s tɛekazt, ad at-semken s tɛekazt. I yɛedan di lgira-nni i yɛedan, i yɛedan, ad d-nawweḍ ayelli taɛzizt-iw ay-jemɛen ama ar tejmaɛt, ama ar yiwet n lḥara n ttdakal mera, nettugad. Irgazen a yelli fuken-ten, fiḥel ad am-d-iniy, fiḥel, irgazen fuken. Surtu ma ddzn-d a Tefrit, d Buɛwen, d At ɛica. Akken ara ten-walin yergazen tteddun-d wid-ak, ttedun-d lɛeskar-agi, wanag lɛeskar n yiṛumiyen; zik d lmal i ykesen aṛṛebiɛ armi d ass mi i llan akka ixabiten ifuken tafart n yirgazenAh; ayu ass-agi ddan-d aTefrit, d Atɛica, d Buɛewen. I qedren deg yirgazen, i sɛewjen deg yimegra n yirgazen msakit. Aha dayen tura si finin. Ah!di tleqqamt n At aɛli iwsawen, ih ! a lexwayar n yimir, d tirebaɛ mera, d tirebaɛ ; d Arezqi n Welḥağ, Latamen At mecqant, setti-im ad ttig Rebbi n ṛṛeḥma, jedi-im, axxam ahaxxam n At winaten, ad yig Rebbi ččan iqaray n sen; uxxam At sɛada, aken kan ticki nettuqrab akka nettuqrab amɣar-nney ad t-yig Rebbi n ṛṛeḥma, akka, At sɛada ttnusun dina yii-d ass ass-agi ad n-nsen At sɛada deg yiger n tqayed jemaɛ

liman ma neka-tt ula d neknni, ad nemger deg yiḍ ayaxir semeḍ lḥal, ad netta, ad ruḥen At sɛada-nni ad d-feken yiwet di teslatin nsen ad arnuɣ nekki neɣ Tasaɛdit Tamusat, ad nruḥ ad nawi imensi ad nens ai lexla ad nemger, deg yiḍ ad nemgar, deg zal ad nestaɛfu imir arraw-is ! Ah! Tewwi-iyi-d iğadarmiyen acu ara m-id hedreɣ ɣef uraw-iw a yelli, nenuɣ, nenuɣ tewwi-iyi-ten-d Tillult. Ruḥ tura mayella win ara as-yinin diri-tt assa. Ɛecrin n ssnin-agi ur iyi-iluɛa Buxalfa a tarwa n tɛebbuḍt. Haca Ferḥat ; sɛiɣ yiwen kan, wamag Buxalfa tiwwi-t Tillult. Ih! Ay wteɣ tagut s aḍu ɣef uraw-iw yeğğa-ten-id d imecṭaṭaḥmerra, sekreɣ-ten-id s uyenat; s lbiɛw cra. Anda im-nniɣ akka wwḍeɣ aɛi?. Wwḍeɣ Zubga, wwḍeɣ At lɛarbi, wwḍeɣ taddert u, Ifarḥunen, wwḍeɣ abrid Gnnaris ; ak, ak yelan d tamurt merra, merra, akka; Illulen merra di lqern almi d lqern. Mḥaga, d At ɛica, Agrsafen, d yiɣil n Bukyasa, d Tifrit Umalek, iy merra, merra timura-agi merra nenuda-tent-id s lbiɛ w cra. Zzit ad idu, lwarq ad idu, ih ur nesɛi baba-tenaɣ, ur nesɛi yema-tenaɣ, yema ad tt-yig Rebbi n ṛṛeḥma tɛewen-itideg uraw-iw. A! ad nekker, nwet tagut s waḍu, tura i inekr-iyi Buxalfa, a iwwiɣ n daɛwesu. <CORPUS> <TRANSCRIPTION PHONETIQUE>:

[nəkwniziɣadnəkaꞧðəgidnssafəgiðanoheaatsəʁʁarieaadnaɣwemðieajjapaddamidneppədane ʃʁojsəkso:minfoksəkso:nnianʕədianənnədjaʁvarminfokjaʁvarnniaʕədianrfeðiqətarəneasəvħ ieannrohʁərʃʁojnetsəwqamamrdijanawodajelliarozəmmoranwiiqʃarəppaʁromðəgʃiwanntaʁ ontsawijarajʕajijagiitswinakkaməddənwollahartsiðətsajellianawiiqʃrannnippaʁromðəgʃiwan ntaʁanawədakənnəmwolahtsislðəjxajaeðəgmaranajiafəldʒðraajamanfokitsidðəgaxəfmaradn arsaʕəqqaadawdaʁaldʒeðraaðxəzraʁeazəmoremaofiʁaʕəqqaareqaʃoʃeaðqəjaʁaeidjəqdaʁħə mjaʁaꞧꞧəzqajəllimatʃameoraaꞧʒʒijntoraadnarsajəllieamədieppasaʁinaʁjazzaanətsdəgiriqʃər annniəpwʁromeʃfadjəmmamatsiɣrəppinəꞧꞧahmasətsimatsyrəppillʒənnəeolaðjemmameəjəħ qədanətsdəgiraʁromnniakaarzzaemanofadaʕəqqaaqoranaenəssəxdəjswəʁromnniurnəsʕiꞧdʒ wazurnəsʕðaʁromnnikanadnawədsjəfaꞧħðaməqranadnafeavvarvoʃeamiqəʃʃadənatsnetʃttsaz iðantsozjintanoħmajfokozəmmornnitsoɣamaeəfokeoɣanniðnqəʃetəvħirintsimeɣꞧiwinðinora rtsiʕelafinəgəzggarənsarwaeenməddənðəgwəddʒivəntsajarzaijxajaeðazəmmorirgazəntsijarzi winnziχdifəllaħənmaꞧꞧaiʕəqlijənaðəmxareajənnnaʕmaakwəðjirənmaraaðmxareajenkolijxirj ətsziðimirkolijxirjətsziðeoraðənkkəzinekkəznjarzaqeoraursinaʁaraðhəðraʁaeifaꞧʁiigʕəddan fəllanaʁalahalahnniʁasɑ:siðireppiurijʁəllijittiʒardənfakwa:rdənfakwa:siðiɑ:rəppiuriʁəllijitti ʒinnaardənfakeirnijinnanətsfarasaðatsʒrəddiʁajəwqamiwasmijəweaʁeaɣoesadojəmoewarg gaziwijxəmsawəejaeinðaməʃtoħjədʒajidꞧəvʕaigwardaneamoreeəllaeamoreeəllɑnħəndikaꞧe ppinʃəkrikeənzaevarqoqeeənzaeerəmantjənzaəjfəjəfəjeənzaðəgmaradnəkkareamoreələfnin ddonienppəditsnəppədzovgaaeəʕəvəllaaeəlʕarvinəppədajəlliifarħonənnəppədkoləʃsjaħmojo karmosðəgmaradnəkkərhorhorhorasskaməjtsiχliwinadnawədanəzənzadnatʃarjaħmojnniðak

arðəjəvsəjðjvaŧaŧaanoħnəznozojsjəðriməneamoreeəllaajəllinxəðəmnəttʃanħəndikaɍəppinʃək rikeorawəllahaaɍəppiarsiniʁðijaʕmarfiʁaʁalaxjaarakkaidofiʁimaniwtsʁimiʁakkagwuxxama nənʁəʃanəzðəmadnawiamanðiləfninddoniemaɍatʃatʃarənteliwaarwasifaranoħdəgiđwəllahar ðaxejtəməðwinidəntʃattʃareivətsijinntsroħodsaxxami:igʕədanihiəlaqwanənziχɑ:ziχeavarvoʃ emaeəttʃidtsaħəqɍəppiarxirnmi:jawmi:jinðaχsomntoraeavarvoʃenziχakkatseimʁarinmawla ʃmaeəfokeəjənatseχwarfatsonəvðoatsijilvəʃnamaefokləvʃnaeavarχantatsiliemələleaðjilliov əlodðaziðanijqoekoləʃðaziðanðawənʕaneoraeimʁarinarssantijqəʃeijməzjintsədontʕarjanəlla hivarəχðajaigəllaneorahaeawiglan

tsiðətsnəznozojlləfejətsəmʁajdlləfənetsawitsʕəvganlləfeimirjəlozlħalʃeaqənməddənlqoead nətʃartsʕəvganlləfearemoraaðaʁdəfkənavəlodtsifraʃovəlođatsidnəzəðseəsireadnəgaraʁrome avarvoʃetsaziðantaʁromisðazieiʃkiolaʃeifraʃnnintsəksədazəgzaweoraazəgzaweoramaeəɣrəŧ idðaʁromwəllahmaemənađaðjəkəsðəgmmiχntsədəzamaqʃoraðeənɣarðaʁromaɍəppiðiemaxj oqeagieaɍwaeađsaearidaʁðiʕəgwonənaʕʕiajaməzoʁimiχəmkanimidhəðraʁakkawanaɣʒʒam miaddhəðraʁmorsinəʁlahðorħəqɍəppimasənaʁajəlliləhðorħaʃaajənjəllankanakkasofəllaeənn ajassaʕədijaeariħantatsiðəkərɍəppisjxireəmətoenləwnisaeʃilatsakənaradawəđənkaneinasajo olaʃlwaxdafəllialamawalaʁeənvərandijsaɍxođennəsənsəddawpəxxamnʒohraTʕəʒoɣeðilħara nninʒohratʕəzoɣewinaofəllaajoɑ:jəsəemaeimaʕzozinolaʃlwaxdafəlliarðwajiʁvrandisərxođə nnsənsəddawwəxxamnʒohratʕəzoɣeadawđənajəllijəxajaeaðsəgəʒgiʒəntðəgəxxamənaeəntəz oʁzoʁorənaʁamzonðjmajwinwaʕankanaesəfðənseməgħəlewinwəʕanaesəfðənseʕəkwazəea ðtssəməχənseʕəkwazəeigʕədanðilgirammiigʕədanadnawəđajəllieaʕzizeiwaʁʒəmʕənamaaree ʒmaʕeamaarjiweelħaranətsðakajmərɍanətsagwaðirgazənajəllifokəntənfiħəlamdiniʁfiħəlirga zənfokənsortomaddandatəfrieðvoʕwənðaeʕiʃaakkənaraeənwalinjargazəntsdondwiðaχtsdon djʕəsəkaragiwanaɣijʕəskariɍomijənziχðjmajigχəsənaɍɍəviʕarmiðassmillanakkaixabieənifok əneafareəgargazənɑhajo:assagiddandatəfrieðaeʕiʃaðvoʕwəniqədrənðəgargazənisʕewʒənðəg məgrađəgargazənmsaχieahðeajəneorasifininahðieləqamtnaeʕjiiwsawənihaləxwajarəgimirts irəbaʕmərɍatsirəbaʕðarəzəqipəlħadʒlatamənaeməʃqansətsimatsiɣaɍəppinɍəħəmaʒədimaxxa mnaewinaeənaðiɣaɍəppittʃaniqarajnsənwəxxamaesəʕaðaakənkaneiʃkinətsoqravakkanətsoqr avamʁarnaʁaeiɣaɍəppinɍəħmaakkaaesʕaðatsnosənðinajinədassagiaðnəsənaesʕaðagiɣarətq ajənðʒəmaʕjimanmanəkkatsojaðnəknianəmɣardəgiđajaxirsəməđəjħajaðnətsaaðroħənaesʕað anniadəfkənjiweəðieəsjaeinnsənaðanərnonəkkinəʁeasəʕðieeamosatsanoħanawiimənsianən sðijaxjaanəmɣardəgiđanəmɣarðəgzajanəseaʕfoimiraheeppijididʒadarmijənaʃoaramdhəðraʁ ʁafarrawiwajəllinənoʁnənoʁeəppijieənideillolerohʕeoramajəllawarasjininðiritsassaʕəʃrinləs najagiorjiloʕavoxajfaaearwatəʕəbotħaʃafarħaesʕiʁjiwənkanwanaɣvoxajfaeəppieeilloleihaje weaʁeaɣoesađoʁafarawiwjədʒajiniddiməʃtataħmərɍasəkraʁnidsojənaesjviʕwəʃraandaməni ʁakkapđaʁaʕipəđaʁzovgapəđaʁaejʕarvipəđaʁeaddareifarħonənpəđaʁavriðgnnarisχaχajəla

ntsmorəeməꞧꞧaməꞧꞧaakkaillojənməꞧꞧaðilqarənajamaðjqarənmħaɣaðaeʕiʃaaɣwrsafənðjiʁijn voχjasatsifrieomajəχijməꞧꞧaməꞧꞧaeimorajagiməꞧꞧannoðaeəntsjviʕwəʃrazieaðjədojwarqaej ədoihornəsʕivavaenaʁornəsʕijəmaenəʁjəmaatsiɣarəppinꞧꞧəħmaeʕawənijiðəgarrawiwɑadnə karnəwəeeaɣwoeswaɗoeoraijnəkərijivoxajfaajppaʁndaʕwəssu]

\- <Métadonnées>

<T>**traditions et coutumes**</T>

<L>**Village: Sahel/ Commune: Bouzeguène / Daira: Bouzeguène/ Wilaya: Tizi-ouzou**</L>

<D>**novembre 2007**</D>

<Dial>**langue kabyle**</Dial>

<Tr>**I- Notation usuelle, II- Transcription phonétique**</Tr>

<Enreg>**méthode semi-directive / caméscope (audiovisuel)**</Enreg>

<Loc>**Nna ouardia/ sexe : Féminin/ âge : 79 ans/ monolingue**</Loc>

<Enq>**Karima HABBI**</Enq>

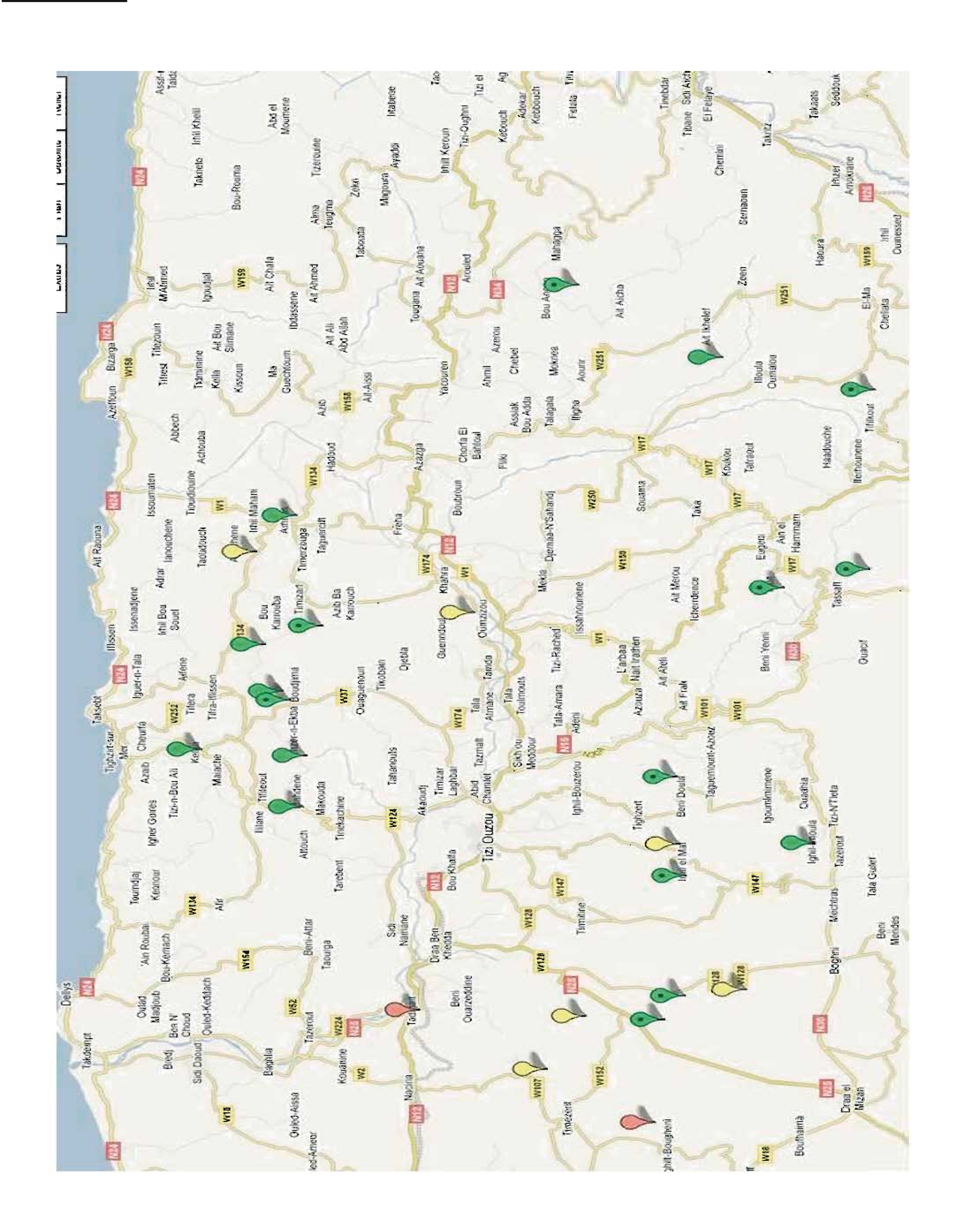
<FichVideo href="c:\..." />

<Rem />

</Métadonnées>

</CORPUS>

Annexe 4:



Annexe 5 :

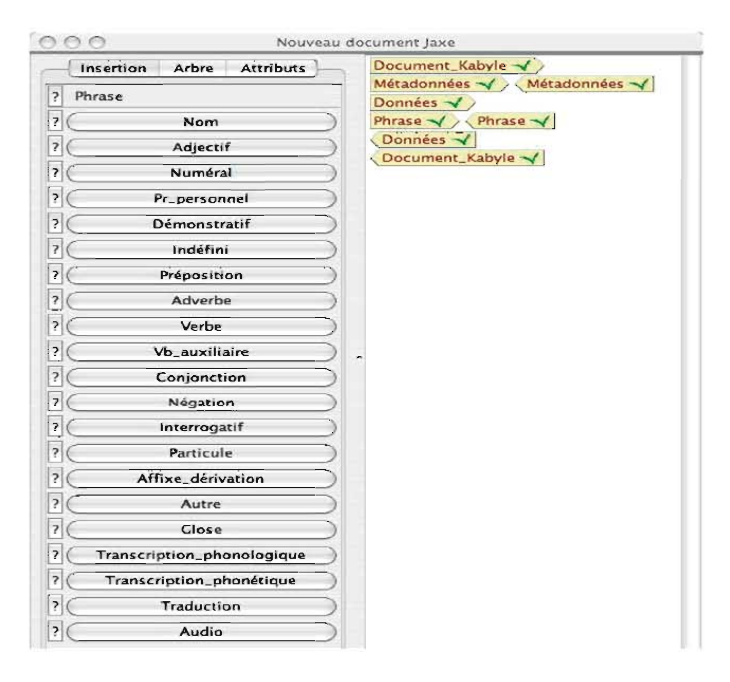

```
<?xml version="1.0" encoding="ISO-8859-1"?><JAXECFG>
<DESCRIPTION>Configuration pour corpus de kabyle -
UMMTO-UNIL</DESCRIPTION>
<RACINE>
<BALISE nom="CORPUS" titre="Référence du corpus"
type="division">
<TEXTE/>
<SOUSBALISE nom="Métadonnées"/>
<SOUSBALISE nom="Données"/>
</BALISE>
</RACINE>
<MENU titre="Métadonnées">
<BALISE nom="Métadonnées" titre="Informations sur ce
corpus" type="division">
<TEXTE/>
<SOUSBALISE nom="T"/>
<SOUSBALISE nom="L"/>
<SOUSBALISE nom="D"/>
<SOUSBALISE nom="Dial"/>
<SOUSBALISE nom="Tr"/>
<SOUSBALISE nom="Enreg"/>
<SOUSBALISE nom="Loc"/>
<SOUSBALISE nom="Enq"/>
<SOUSBALISE nom="FichSon"/>
<SOUSBALISE nom="FichCarte"/>
<SOUSBALISE nom="FichPhoto"/>
<SOUSBALISE nom="FichVideo"/>
<SOUSBALISE nom="Rem"/>
</BALISE>
<BALISE nom="T" titre="Titre" type="string">
<TEXTE/>
</BALISE>
<BALISE nom="L" titre="Lieu" type="string">
<TEXTE/>
</BALISE>
<BALISE nom="D" titre="Date" type="string">
<TEXTE/>
</BALISE>
<BALISE nom="Dial" titre="Dialecte(s)" type="string">
<TEXTE/>
</BALISE>
<BALISE nom="Tr" titre="TypeTranscription"
type="string">
<TEXTE/>
</BALISE>
```

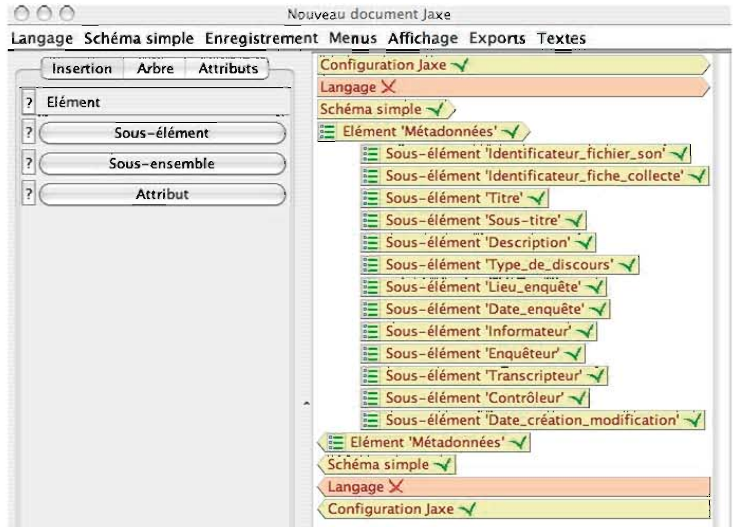
```
<BALISE nom="Enreg" titre="InfosEnregistrement"
type="string">
<TEXTE/>
</BALISE>
<BALISE nom="Loc" titre="Informateur(s)" type="zone">
<TEXTE/>
</BALISE>
<BALISE nom="Enq" titre="EnquÍteur" type="string">
<TEXTE/>
</BALISE>
<BALISE nom="FichSon" titre="Son" type="zone">
<ATTRIBUT nom="href" presence="obligatoire"/>
<TEXTE/>
<ATTRIBUT nom="start" presence="optionelle"/>
<ATTRIBUT nom="stop" presence="optionelle"/>
</BALISE>
<BALISE nom="FichCarte" titre="Carte" type="string">
<ATTRIBUT nom="href" presence="obligatoire"/>
<TEXTE/>
</BALISE>
<BALISE nom="FichPhoto" titre="Photo" type="string">
<ATTRIBUT nom="href" presence="obligatoire"/>
<TEXTE/>
</BALISE>
<BALISE nom="FichVideo" titre="Vidéo" type="string">
<ATTRIBUT nom="href" presence="obligatoire"/>
<TEXTE/>
</BALISE>
<BALISE nom="Rem" titre="Remarques" type="zone">
<TEXTE/>
</BALISE>
</MENU>
<MENU titre="Données">
<BALISE nom="Données" titre="Texte transcrit"
type="division">
<TEXTE/>
<SOUSBALISE nom="Enoncé"/>
<SOUSBALISE nom="Phrase"/>
<SOUSBALISE nom="Mot"/>
<SOUSBALISE nom="Monème"/>
<SOUSBALISE nom="Trad"/>
<SOUSBALISE nom="Phono"/>
<SOUSBALISE nom="Phonet"/>
<SOUSBALISE nom="Graf"/>
<SOUSBALISE nom="ChLg"/>
</BALISE>
<BALISE nom="ChLg" titre="Autre langue" type="string">
<ATTRIBUT nom="code_langue" presence="obligatoire"/>
```

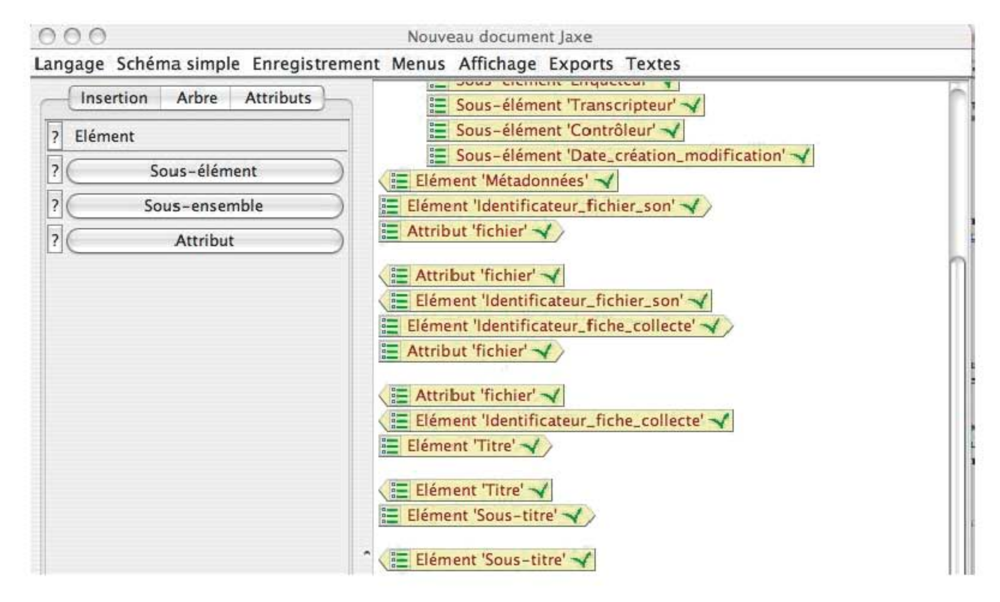
```
<TEXTE/>
</BALISE>
<BALISE nom="Enoncé" titre="Enoncé" type="string">
<ATTRIBUT nom="id" presence="obligatoire"/>
<SOUSBALISE nom="Phrase"/>
<SOUSBALISE nom="Mot"/>
<SOUSBALISE nom="Monème"/>
<SOUSBALISE nom="Trad"/>
<SOUSBALISE nom="Phono"/>
<SOUSBALISE nom="Phonet"/>
<SOUSBALISE nom="Graf"/>
<TEXTE/>
</BALISE>
<BALISE nom="Phrase" titre="Phrase" type="string">
<ATTRIBUT nom="id" presence="obligatoire"/>
<SOUSBALISE nom="Enoncé"/>
<SOUSBALISE nom="Mot"/>
<SOUSBALISE nom="Monème"/>
<SOUSBALISE nom="Trad"/>
<SOUSBALISE nom="Phono"/>
<SOUSBALISE nom="Phonet"/>
<SOUSBALISE nom="Graf"/>
<TEXTE/>
</BALISE>
<BALISE nom="Mot" titre="Mot" type="string">
<TEXTE/>
<SOUSBALISE nom="Monème"/>
<SOUSBALISE nom="Trad"/>
<SOUSBALISE nom="Phono"/>
<SOUSBALISE nom="Phonet"/>
<SOUSBALISE nom="Graf"/>
</BALISE>
<BALISE nom="Monème" titre="Monème" type="string">
<TEXTE/>
<SOUSBALISE nom="Trad"/>
<SOUSBALISE nom="Phono"/>
<SOUSBALISE nom="Phonet"/>
<SOUSBALISE nom="Graf"/>
</BALISE>
<BALISE nom="Trad" titre="Traduction" type="string">
<TEXTE/>
</BALISE>
<BALISE nom="Phono" titre="Phonologie" type="string">
<TEXTE/>
</BALISE>
<BALISE nom="Phonet" titre="Phonétique" type="string">
<TEXTE/>
</BALISE>
```

```
<BALISE nom="Graf" titre="Graphie" type="string">
<ATTRIBUT nom="systeme" presence="optionelle"/>
<TEXTE/>
</BALISE> </MENU> </JAXECFG>
```

Annexe 6, 7, 8 et 9 :






Nouveau document Jaxe
Langage Schéma simple Enregistrement Menus Affichage Exports Textes
Insertion Arbre Attributs
Elément
Sous-élément
Sous-ensemble
Attribut
Sous-élément 'Transcripteur'
Sous-élément 'Contrôleur'
Sous-élément 'Date_création_modification'
Elément 'Métadonnées'
Elément 'Identificateur_fichier_son'
Attribut 'fichier'
Attribut 'fichier'
Elément 'Identificateur_fichier_son'
Elément 'Identificateur_fiche_collecte'
Attribut 'fichier'
Attribut 'fichier'
Elément 'Identificateur_fiche_collecte'
Elément 'Titre'
Elément 'Titre'
Elément 'Sous-titre'
Elément 'Sous-titre'


Nouveau document Jaxe
Insertion Arbre Attributs
Métadonnées
Identificateur_fichier_son
Identificateur_fiche_collecte
Titre
Sous-titre
Description
Type_de_discours
Lieu_enquête
Date_enquête
Informateur
Enquêteur
Transcripteur
Contrôleur
Date_création_modification
Document_Kabyle
Métadonnées
Métadonnées
Document_Kabyle

Annexe 10 :

Le corpus ségmenté :

- <Bruit de fond> (00 : 00 s. – 00 : 03 s.)

B- ‖ Ad wen-d-hedreɣ | ɣ ef ***cinquante-huit*** ? ‖

SPV synt. Prépos.

A- ‖ Ah ? ‖

- <Bruit de fond> (00 : 06 s. – 00 : 08 s.)

A- ‖ **<euh : >** ***alors*** **<euh : >**

‖ Aqlaɣ | di Tesga-Mellul | d ssebt |

Présentatif syntagme nominal syntagme nominal

tnayen-u¯ecrin ɣ uct | ttes¯a ɣ ir ṛṛbe¯ ‖

syntagme nominal syntagme nominal

C- ‖ ɣ ir ṛṛbe¯ ‖

syntagme nominal

B- ‖ Aqlaɣ | di Tesga-Mellul | ttes¯a-wa:crin… |

Présentatif syntagme nominal ?

d ssebt | ttes¯a-wa:crin di ɣ uct |

syntagme nominal syntagme nominal

ssa¯a | attan | d tes¯a **<u: (hésitation)>**

indicateur de thème présentatif syntagme nominal

ɣ ir ṛṛbe¯ ‖

syntagme nominal

- <Bruit de fond> (00 : 26 s. – 31 : 08 s.)

- ‖ Di ***la pression*** n tmanya-uxemsin | yella | lliɣ |

Synt. prépos auxiliaire SPV

di Micli ‖

synt. prépos.

- ‖ Ass-nni | <amar n imjuhad> ‖

Autonome syntagme nominal

- ǁ Ilaq ǀ ad n*regroupi* ǀ ɣer tudrin ǁ
SPV Synt.Prépos.

- ǁ Nekkenni ǁ

- ǁ Iṣubb-iyi-d ǀ Dda Sliman ǀ bessif ǀ
SPV expansion référentielle adverbe

si Micli ǁ
Synt.Prépos.

- ǁ Nṣubb-d ǀ nebbweḍ ǀ ɣer ssbiṭar ǀ
SPV SPV synt. Prép.

nufa-d ǀ *l'embuscade* ǁ
SPV expansion directe

- ǁ Axaṭar ǀ nʿeṭṭel ǀ di micli ǁ
Subordonnant SPV synt. Prép.

yeɣli-d ǀ ṭṭlam ǁ.
SPV expansion référentielle

- ǁ Netta ǀ d *couvre-feu* ǀ
Pronom personnel indépendant Syntagme nominal

amek ǀ teɣli ǀ *l'embuscade* amezwaru ǁ
interrogatif SPV expansion référentielle

- ǁ Neṛğa ǀ ixeddamen ǀ s ukamyun ǀ mi d-ffɣen ǀ
SPV expansion directe expansion indirecte
Proposition 1

axaṭer ǀ ur nezmir ara ǀ ad nʿeddi ǁ
subordonnant Proposition 2

- ǁ n*traversi* ǀ ɣer lğiha-agi ǁ.
SPV synt. Prép.

- ǁ nniɣ-as ǀ ma yella ǀ nufa ǀ *l'embuscade* ǀ
SPV subordonnant auxiliaire SPV expansion directe

ɣer zdat ǀ ma ʿeddan-d ǀ ad ɣenɣen ǁ
synt. Prép. subordonnant SPV (proposition1) (prop.2)

- ǁ <nebbwi-ten | daxel> | yeṭṭerḍeq | ukamyun |

SPV adverbe SPV expansion référentielle

neqqim | d ***cca¯b*** | ger-aneʏ |

SPV cordonnant synt. Prép.

en dehors | ***des personnes de <pipa>*** ǁ

??????????????????????????????

- ǁ int-as | i ***la supérieure*** | ***i la mère*** ǁ

SPV Expansion indirecte Expansion indirecte

- ǁ nnan-as | ğğan-aʏ | weḥd-nneʏ | di ssbiṭar |

SPV SPV synt. Prép. synt. Prép.

neṭṭes | ***en chirurgie*** ǁ

SPV synt. Prép.

- ǁ Akken | d ttnac | n deggiḍ |

Adverbe syntagme nominal synt. Prép.

qel¯en | ʏ er tewrirt | s tsita | s i¯ejmiyen ǁ

SPV synt. Prép. synt. Prép. synt. prép.

- ǁ Ass-nni | wwten | Wizan n Muḥend Wa¯meṛ ǁ

Autonome SPV expansion directe

- ǁ huzzen-tt | deg ufus ǁ

SPV synt. Prép.

- ǁ nʏ an-as | taqcict | ʏ ef yiʏ il-is ǁ

SPV expansion directe synt. Prép.

- ǁ nʏ an | aεejmi ǁ

SPV expansion directe

- ǁ nʏ an | alews-is | Sa¯id At Ṭṭaleb |

SPV expansion directe expansion directe

dʏ a | wwintt-id | ʏ er ssbiṭar ǁ

connecteur (autonome spécifique) SPV synt. Prép.

- ǁ Amek | i s-xedmen ǁ

Interrogatif relatif (proposition relative)

- ǁ Ssawlen ǁ

SPV

- ǁ…- **<Bruit de fond>** ….

- ǁ rrfed-itt ǀ imir ǀ si Micli ǀ ɣ er Tizi-Wezzu ǁ

SPV adverbe synt. Prép. synt. Prép.

- ✓ // Azekka-nni, / Ɛefsen-aneɣ / aƐwin d yiṣurdiyen //
  Adverbe SPV expansion directe
- ✓ // nnan-as / i baba lḥağ, i baba lḥağ Muḥend / kker //
  SPV expansion indirect SPV
- ✓ // yeɣ li / gar tṣeddarin // ;
  SPV synt. Prép.
- ✓ // yenna-as / lukan d lƐibad / i aɣ -d-wwin //
  SPV subordonnant syntagme nominal proposition relative
- ✓ // ur yettsuɣ u ara / s yiṣurdiyen …**la salle d'eau /** d axxam piṭru //;;
  SPV synt. Prép. …. Syntagme nominal
- ✓ // akken ajirikan n waman, acifun ma jajin iɣ min iɣ amen yiṭu i aɣ -ttawin. //
- ✓ // Wwin qbel / deg yimeqqranen // ,
  SPV adverbe synt. Prép.
- ✓ // wwin / CaƐban n WaƐliqa / ad fell-as yeƐfu Rebbi //
  SPV expansion directe
- ✓ // rnan / Ɛmara iqa //
  SPV expansion directe
- ✓ // ad ak-rnun Muḥend WaƐ //
  SPV expansion directe
- ✓ // meqqar … Rnan / dadda-k ṣalaḥ, ṣalaḥ At-AƐli //
  SPV expansion directe
- ✓ // i walbeƐḍ-nneɣ , //
  synt. Prép.
- ✓ // wwin / Yunes At-SaƐdi //

SPV expansion directe

✓ // ula d netta / ur yeqqar ara //

SnV SV

✓ // d baba-s / i yeqqaren / deg-sen / di Ɛmara, //

Syntagme nominal proposition relative synt. prép. Synt. prép.

✓ // uɣ alen-d / ɣ ur-i / ur qqareɣ ara, //

SPV Synt.prép. SPV

✓ // daɣ en Ferḥat Sliman akked Muḥend At-WeƐli –Ulḥusin /

Adverbe nominal cordonnant nominal

✓ // ur qqaren ara irkel //

SPV adverbe

✓ // **mais** mbeƐd …kan xemsa n taddart / ufan / xemsa **limumbr**. //

Cordonnant adverbe synt.nominal SPV expansion directe

✓ // Azekka-nni, ad d-nruḥ / ar taddart //

Adverbe SPV synt.prép.

✓ // sukken-aɣ -d / seg Furbiyen, / dinna s-ddaw Uwrir // ,

SPV Synt.prép. déictique synt. Prépos.

✓ // Dda Muḥend-Ḥemmu fell-as yeƐfu Rebbi /

Indicateur de theme

//yeƐwej-as uttbadri //

SPV Expansion référentielle

✓ // deg uzɣ al **hi d-nemlal li-d nhar** / s ukubri-nsen, / s **les para** //

✓ //…lḥara n xali-k Ibrahim … / ḍerfen-d / xemsa-nni n taddart //

Syntagme nominal SPV expansion directe

✓ // ḥebsen-ten / ar lhiḍ / miṭrayin-ten /…..nɣ an-ten //

SPV synt.Prépos. SPV SPV

✓ // iruḥ ciṭ-nni ……//

SPV expansion directe

✓ **// après** weḥd-s yerra-t / **un peu de grace** …..//

Adverbe SPV

✓ // itekka-as / s yiwet / s **la balle** / s aqerru // ,

SPV synt. Prép. synt. Prép. synt. Prép.

- ✓ // nejbed-as …// // nenṭel-it ……// // tƐedda ddeƐwa //

SPV SPV SPV expansion référentielle

- ✓ // …….uɣ alen-d **une deuxième fois** //

SPV

- ✓ // nekkni / nfaq // ,

Indicateur de thème SPV

- ✓ // ur d-ufin / …… yiwen / yiwen (**répétition**) / di taddart //,

SPV ????????? synt. Prépos.

- ✓ **[passage d'un autre informateur] / meẓẓi neɣ meqqer //**
- ✓ // nekk / sƐiɣ tafunast //,

Indicateur de thème SPV expansion directe

- ✓ // wwiɣ tafunast / ksiɣ -tt //

SPV expansion directe SPV

- ✓ //…. Deg uxxam / iruḥ lƐesker //;

Synt.prépos. SPV expansion référentielle

- ✓ // …daɣ nekkini / ma ur iruḥ ara lƐesker //

Subordonnant SPV expansion référentielle

- ✓ // tenna-ak tmeṭṭut / dɣ a ad tawiḍ tafunast //

SPV expansion référentielle SPV expansion directe

- ✓ // nekk / ad rreɣ kan / syagi … //

Indicateur de theme SPV adverbe déterminant autonome

- ✓ // yuɣ al, / arraw-is / bdan / lǧemƐa //

auxiliaire indicateur de thème SPV adverbe

- ✓ // ass-nnikat d ssebt / dɣ a i neddukkel //

Adverbe syntagme nominal connecteur

- ✓ // ama d wigad i xeddmen /

Fonctionnel propositionnel SnV relatif prédicatoide

Wigad ur nxeddem ara; //

substitut non personnel prédicatoide

✓ // yiwen n yiḍ, / tettṣubbu tmeṭṭut //
Adverbe SPV expansion référentielle
// tenna-ak / ad awiɣ tafunast //
SPV SPV expansion directe

✓ // nniɣ -as / tura ad nens ……. : //
SPV adverbe SPV

✓ // nuqem ttiƐad / ad nemlil / deg yiɣ zer-nni Bu-Sliman //
SPV expansion directe SPV expansion indirecte

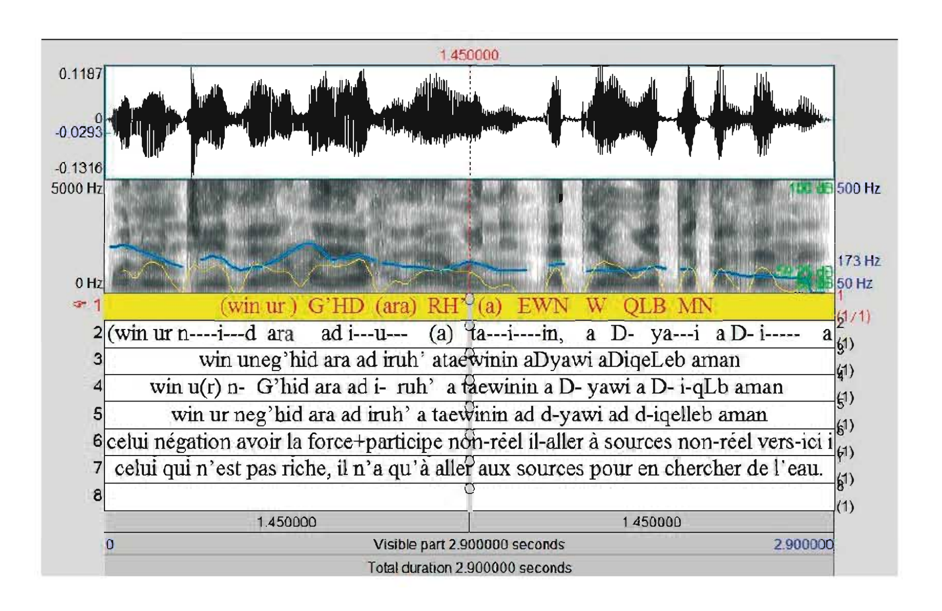
✓ // … Xedmen-asen ratissage //
SPV expansion indirecte expansion directe

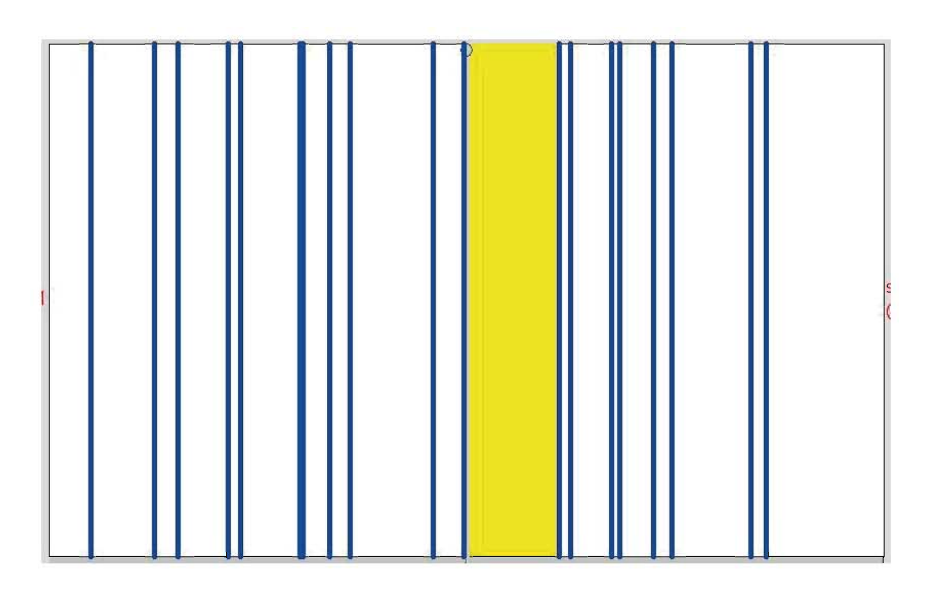
✓ // i ṛemḍan-nni / deg yiḍ, nerğa / armi d sebƐa //
Syntagme nominal Synt.prépos. SPV syntagme prédicatoide

✓ // xedƐen-aɣ akk //
SPV adverbe

✓ // qqimeɣ -d / ala weḥd-i / di teswiƐt-nni / n Sid-Lḥusin //
SPV synthème adverbial synt. prépos. Synt.prépos.

Annexe 11 et 12 :





Building an annotated corpus for Amazighe

Mohamed Outahajala[1], Lahbib Zenkouar[2], Paolo Rosso[3]
[1] Royal Institut for Amazighe Culture, Rabat, Morocco
outahajala@ircam.ma
[2]Ecole Mohammadia d'Ingénieurs, Rabat, Morocco
zenkouar@emi.ac.ma
[3]Natural Language Engineering Lab - EliRF, DSIC, Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es

Abstract

This paper gives an overview of the morpho-syntactic features of the Amazighe language and corpus encoding, afterwards we present our experience of constructing an annotated corpus with part-of-speech (POS) information. The annotated corpora consist of 20,667 Moroccan Amazighe tokens chosen from different materials; it is to our knowledge the first one dealing with Amazighe language. The experience is also meant to give a handle on the encoding and tagging processes of the aforementioned corpus.

1. Introduction

Amazighe language is spoken in Morocco, Algeria, Tunisia, Libya, and Siwa (an Egyptian Oasis); it is also spoken by many other communities in parts of Niger and Mali. It is a composite of dialects of which none has been considered as the national standard in any of the already mentioned countries. With the emergence of an increasing sense of identity, Amazighe speakers would very much like to see their language and culture rich and developed. To achieve such a goal, some Maghreb states have created specialized institutions, such as the Royal Institute for Amazighe Culture (IRCAM, henceforth) in Morocco and the High Commission for Amazighe in Algeria. In Morocco, Amazighe has been introduced in mass media and in the educational system in collaboration with relevant ministries. Accordingly, a new Amazighe television channel was launched in first March 2010 and it has become common practice to find Amazighe taught in various Moroccan schools as a subject.

Over the last 8 years of its creation, IRCAM has published more than 150 books related to the Amazighe language and culture, a number which exceeds the whole amount of Amazighe publications in the 20th century, showing the importance of

an institution such as IRCAM. However, in Natural Language Processing (NLP) terms, Amazighe, like most non-European languages, still suffers from the scarcity of language processing tools and resources. In line with this, and since corpora constitute the basis for human language technology research; yet they are difficult to have for a number of languages - particularly annotated ones. In this paper we try to shed light on an experience of constructing an annotated corpus along the information provided by part-of-speech (POS); the corpus consists of over than 20k Moroccan Amazighe tokens. The experience is also meant to give a handle on the encoding and tagging processes of the aforementioned corpus. To our knowledge the annotated corpus presented in this paper is the first one to deal with Amazighe. This resource even though small, is very useful for training taggers, themselves basic tools for more advanced NLP.

The rest of the paper is structured as follows: in Section 2 we present an overview of the Amazighe morpho-syntactic features. Then, in Section 3 we describe corpus encoding. In Section 4 we present the manner in which the annotation was undertaken. Finally, in Section 5 we draw some conclusions and describe the work to be done in the future.

2. Morpho-syntactic specifications and tagset

2.1. Some Amazighe language features

Amazighe belongs to the Hamito-Semitic/"Afro-Asiatic" languages (Cohen 2007) with a rich morphology (Chafiq 1991, Boukhris et al. 2008). Amazighe is used by tens of millions of people in North Africa mainly for oral communication. According to the last governmental population census of 2004, the Amazighe language is spoken by some 28% of the Moroccan population (millions).

Amazighe standardization is taking into consideration its linguistic diversity. As far as the alphabet is concerned by standardization, and because of historical and cultural reasons, Tifinaghe has become the official graphic system for writing Amazighe. IRCAM kept only pertinent phonemes for Tamazight, so the number of the alphabetical phonetic entities is 33, but Unicode codes only 31 letters plus a modifier letter to form the two phonetic entities: ⴳⵯ(g^w) and ⴽⵯ(k^w). The whole range of Tifinaghe letters is subdivided into four subsets: the letters used by IRCAM, an extended set used also by IRCAM, other neo-tifinaghe letters in use and some attested modern Touareg letters. The number reaches 55 characters (Zenkouar 2008, Andries 2008). In order to rank strings and to create keyboard

layouts for Amazigh in accordance with international standards, two other standards have been adapted (Outahajala and Zenkouar, 2008):

- ISO/IEC14651 standard related to international string ordering and comparison method for comparing character strings and description of the common template tailorable ordering;

- Part 1: general principles governing keyboard layouts of the standard ISO/IEC 9995 related to keyboard layouts for text and office systems.

The graphic rules for Amazighe words are set out as follows (Ameur et al 2006a, 2006b):

- Nouns, quality names (adjectives), verbs, pronouns, adverbs, prepositions, focalizers, interjections, conjunctions, pronouns, particles and determinants consist of a single word occurring between blank spaces or punctuation marks. However, if a preposition or a parental noun is followed by a pronoun, both the preposition/parental noun and the following pronoun make a single whitespace-delimited string. For example: ⵖⵔ (ɣr) "to, at" + ⵉ (i) "me (personal pronoun)" results into ⵖⴰⵔⵉ/ⵖⵓⵔⵉ (ɣari/ɣuri) "to me, at me, with me".

- Amazighe punctuation marks are similar to the punctuation marks adopted internationally and have the same functions. Capital letters, nonetheless, do not occur neither at the beginning of sentences nor at the initial letters of proper names.

The English linguistic terminology used in this paper was extracted form (Boumalk and Naït-Zerrad, 2009).

2.2. Amazighe tagset

Based on the Amazighe language features presented above, Amazighe tagset may be viewed to contain 13 parts-of-speech with two common attributes to each one: "wd" for "word" and "lem" for "lemma", whose values depend on the lexical item they accompany.
The defined Amazighe elements and their attributes are set out in what follows:

POS	attributes and subattributes with number of values
Noun	gender(3), number(3), state(2), derivative(2), POSsubclassification(4), person(3), possessornum(3), possessorgen(3)
Adjective/ name	gender(3), number(3), state(2), derivative(2), POS subclassification(3)

of quality	
Verb	gender(3), number(3), aspect(3), negative(2), form(2), derivative(2), voice(2)
Pronoun	gender(3), number(3), POS subclassification(7), deictic(3), person(3)
Determiner	gender(3), number(3), POS subclassification(11), deictic(3)
Adverb	POS subclassification(6)
Preposition	gender(3), number(3), person(3), possessornum(3), possessorgen(3)
Conjunction	POS subclassification(2)
Interjection	
Particle	POS subclassification(7)
Focus	
Residual	POS subclassification(5), gender(3), number(3)
Punctuation	punctuation mark type(16)

Table 1. A synopsis of the features of the Amazighe POS tagset with their attributes and values

In Table 1, the subcategories of the noun are:

(i) Gender: 1. Masculine	2. Feminine	3.Neuter	
(ii) Number: 1. Common	2. Singular	3. plural	
(iii) Derivative: 1. No	2.Yes		
(iv) POS type: 1. Commun	2. Numeral	3. Parental	4. Proper
(v) State : 1. Construct 2.Free			

When a noun is parental, it it might have 3 aditional attributes: possessor gender, possessor number and possessor person.Adjectives, called also quality names inherit the propreties of nouns. It may be a derivative or not. The subcategories of the adjectives are:

POS type : 1. Ordinal	2. Qualificative	3. Relational

The subcategories of pronouns are:

1. Demonstrative	2.Exclamative	3.Indefinite
4.Interrogative 5. Personal	6. Possessive	7. Relative

The verb attributes that have been used in our tagset are:

(i)	Gender: 1. Masculine	2. Feminine	3.Neuter
(ii)	Number: 1. Common	2. Singular	3. plural
(iii)	Person: 1. 1(first) 2. 2(second)		3. 3(third)
(iv)	Aspect: 1. Aorist	2. Perfective	3.Imperfective.
(v)	Form: 1. Imperative	2. Participle	
(vi)	Derivative: 1. No	2.Yes	
(vii)	Voice: 1. Active	2. Passive	

Aspect attribute have "negative" as sub attribute with two values: negative and positive, when the aspect is equal to perfective. The subcategories of determinants are:

1. Article	2. Demonstrative	3.Exclamative	4.Indefinite
5. Interrogative	6. Numeral	7. Ordinal	8. Possessive
9. Presentative	10. quantifier	11. other	

The adverb types are subdivided into:

1. Interrogative	2. Manner	3. Place	4. Quantity
5. Time	6. Other		

The subcategories of particles are:

1. Interrogative	2. Negative	3. Orientation	4. Predicate
5. Preverbal	6. Vocative	7. Other	

Residual label stands for attributes like currency, number, date, mathematical marks and other unknown residual words. The punctuation category contains all punctuation symbols, such as (?, !, :, ;, .). Elsewhere, conjunctions are subcategorized to coordination and subordination conjunction.

3. Corpus encoding

3.4 Writing systems

Amazighe corpora produced up to now are written on the basis of different writing systems, most of them use Tifinaghe-IRCAM (Tifinaghe-IRCAM makes use of

Tifinaghe glyphs but Latin characters) and Tifinaghe Unicode. It is important to say that the texts written in Tifinaghe Unicode are increasingly used.

Even though, we have decided to use a specific writing system based on ASCII characters for technical raisons (Outahajala et al. 2010).

Correspondences between the different writing systems and transliteration correspondences are shown in Table 2.

Tifinaghe Unicode		Transliteration		Used characters in Tifinaghe IRCAM		Chosen characters for tagging
Code	Character	Latin	Arabic	characters	codes	
U+2D30	ⴰ	a	ا	A, a	65, 97	a
U+2D31	ⴱ	b	ب	B, b	66, 98	b
U+2D33	ⴳ	g	گ	G, g	71, 103	g
U+2D33 & U+2D6F	ⴳⵯ	g^w	گ' +	Å, å	197, 229	g°
U+2D37	ⴷ	d	د	D, d	68, 100	d
U+2D39	ⴹ	ḍ	ض	Ä, ä	196, 228	D
U+2D3B	ⴻ	e[1]	يي	E, e	69, 101	e
U+2D3C	ⴼ	f	ف	F, f	70, 102	f
U+2D3D	ⴽ	k	ک	K, k	75, 107	k
U+2D3D & U+2D6F	ⴽⵯ	k^w	گ' +	Æ, æ	198, 230	k
U+2D40	ⵀ	h	ه	H, h	72,104	h
U+2D40	ⵃ	ḥ	ح	P, p	80,112	H
U+2D44	ⵄ	ɛ	ع	O, o	79, 111	E

[1] note : different use in the IPA which uses the letter ə

U+2D45	ⵅ	x	خ	X, x	88, 120	x
U+2D47	ⵇ	q	ق	Q, q	81, 113	q
U+2D49	ⵉ	i	ي	I, i	73, 105	i
U+2D4A	ⵊ	j	ج	J, j	74, 106	j
U+2D4D	ⵍ	l	ل	L, l	76, 108	l
U+2D4E	ⵎ	m	م	M, m	77, 109	m
U+2D4F	ⵏ	n	ن	N, n	78, 110	n
U+2D53	ⵓ	u	و	W, w	87, 119	u
U+2D54	ⵔ	r	ر	R, r	82, 114	r
U+2D55	ⵕ	ṛ	□	Ë, ë	203, 235	R
U+2D56	ⵖ	ɣ	غ	V, v	86, 118	G
U+2D59	ⵙ	s	س	S, s	83, 115	s
U+2D5A	ⵚ	ṣ	ص	Ã, ã	195, 227	S
U+2D5B	ⵛ	c	ش	C, c	67, 99	c
U+2D5C	ⵜ	t	ت	T, t	84, 116	t
U+2D5F	ⵟ	ṭ	ط	Ï, ï	207, 239	T
U+2D61	ⵡ	w	ۋ	W, w	87, 119	w
U+2D62	ⵢ	^ + ي	ي	Y, y	89, 121	y
U+2D63	ⵣ	z	ز	Z, z	90, 122	z
U+2D65	ⵥ	ẓ	ژ	Ç, ç	199, 231	Z
U+2D6F	ⵯ	ʷ	،	No correspondant in Tifinaghe-IRCAM		°

Table 2. The mapping from existing writing systems and the chosen writing system.

A transliteration tool was built, Figure 1, in order to handle transliteration to and from the chosen writing system and to correct some elements such as the character "^" which exists in some texts due to input errors in entering some Tifinaghe

letters. So the sentence portion "ⴰⵙⵙ ⵏ ⵜⵎⴳⵔⴰ" using Tifinaghe Unicode or "ass n tm^vra" using Tifinaghe-IRCAM and with "^" input error will be transliterated as "ass n tmGra" ("When the day of the wedding arrives").

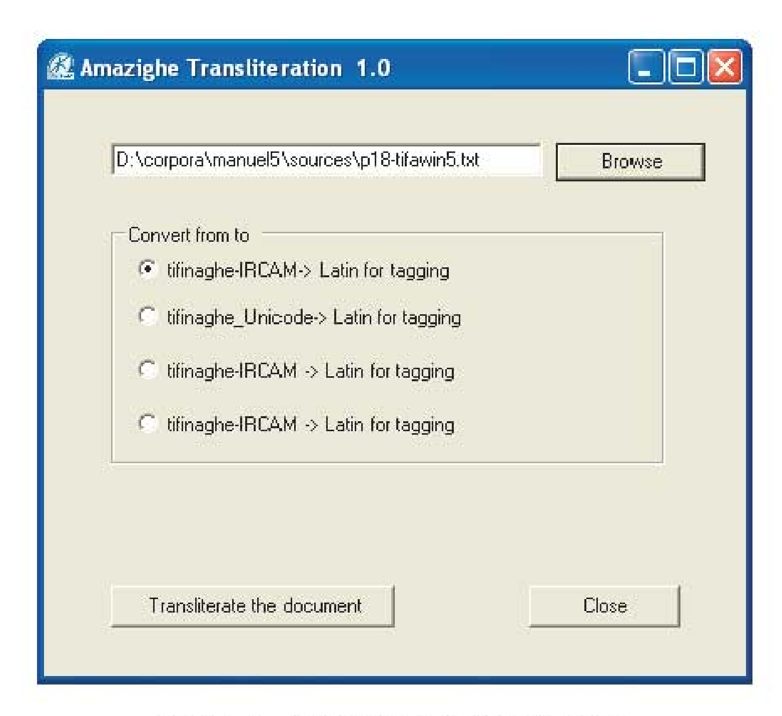


Figure 1. Amazighe transliteration tool

3.5 Corpus description

To constitute our corpora, we have chosen a list of texts extracted from a variety of sources such as: the Amazighe version of IRCAM's web site[2], the periodical "Inghmisn n usinag[3]" (IRCAM newsletter) and three of the primary school textbooks. Table 3 gives a description of chosen sources.

[2] www.ircam.ma

[3] Freely downloadable from http://www.ircam.ma/amz/index.php?soc=bulle

Corpus description	Tokens number	Sentences number
Textbook manual 2	5079	372
Textbook manual 5	2319	179
Textbook manual 6	3773	253
IRCAM web site	4258	185
Inghmisn (IRCAM newsletter)	4636	415
Miscellaneous	602	34
Total	20667	1438

Table 3. Corpus description.

Labeled class	Designation	Occurrences
v	Verb	3190
n	Noun	4993
a	Quality name/Adjective	503
ad	Adverb	516
c	Conjunction	834
d	Determinant	1076
s	Preposition	2775
foc	Focalizer mechanism	91
i	Interjection	40
p	Pronoun	1496
pr	Particle	1593
r	Residual (foreign, number, date, currency, mathematical and other)	178
f	Punctuation	3382
	Total	**20667**

Table 4. Part-of-speech occurrences

After transliterating to the chosen writing system, the corpora, as well as the morpho-syntactic specifications, are encoded using XML. Each token is labeled with the attributes and the sub attributes presented in Table1 using the annotating tool presented below.

We were able to tag 20,667 tokens with a total number of 1,438 sentences. Table 3 summarizes the details of the parts-of-speech occurrences of the chosen corpora.

4. Annotating the corpus

The corpora presented in this paper are manually annotated. This manual annotation, which was performed by a team of four annotators, consists of affecting the different morpho-syntactic features to the tokenized Amazighe texts. Technically, manual annotation was done by the AncoraPipe[4] annotation tool which is an Eclipse Plugin. Eclipse is an extendable integrated development environment. With this plugin, all features included in Eclipse are made available for corpus annotation and developing. AncoraPipe is a corpus annotation tool which allows different linguistic levels to be annotated efficiently by (Bertran et al. 2008), since it uses the same format for all stages. AncoraPipe was used in annotating two corpora of 500,000 words each: a Catalan corpus (AnCora-CAT) and a Spanish (AnCora-ESP) one, (Civit & Martí 2004). The annotation tool interface is organized in different panels where data are shown, buttons and menus are available to perform operations on the corpora, such as grouping and splitting. To perform annotation many panels are used: corpora directory tree panel which allows the user to select a file, sentence list panel shows the sentences of a file, sentence tree permitting to the user to see the data of the annotation level together with lemmas and words and annotation panel performing the annotation operations on the tree and annotate its nodes.
The interface is fully customizable to allow different tagsets defined by the user. In line with this, we have defined a specific tagset to annotate Amazighe corpora. The requirements for AnCoraPipe are: Java 1.5 and the Java graphical library SWT. It includes SWT library for Windows XP. In other platforms, this library comes with the Eclipse package or it can be obtained from eclipse web site directory[5].

The input documents have an XML format, allowing representing tree structures. As XML is a wide spread standard, there are many tools available for its analysis,

---

[4] http://clic.ub.edu/ancora/

[5] http://www.eclipse.org/swt/

transformation and management. Figure 2 shows the annotation of a sentence extracted from a text about a wedding ceremony:

"ass n tmGra, illa ma issnwan, illa ma yakkan i inbgiwn ad ssirdn"

[English translation: "When the day of the wedding arrives, some people cook; some other help the guests get their hands washed"]

```
<sentence>
<n gen="m" lem="ass" num="s"  wd="ass"/>
<s wd="n"/>
<n gen="f" lem="tamGra" num="s" state="construct" wd="tmGra"/>
<f punct="comma" wd=","/>
<v aspect="perfective" gen="m" lem="ili" num="s" person="3" wd="illa"/>
<p postype="relative" wd="ma"/>
<v aspect="imperfective" form="participle" lem="ssnw"  wd="issnwan"/>
<f punct="comma" wd=","/>
<v aspect=" perfective" gen="m" lem="ili" num="s" person="3" wd="illa"/>
<p postype="relative" wd="ma"/>
<v aspect="imperfective" form="participle"  lem="fk" wd="yakkan"/>
<s wd="i"/>
<n gen="m" lem="anbgi" num="p" state="construct" wd="inbgiwn"/>
<pr postype="aspect" wd="ad"/>
<v aspect="aorist" gen="m" lem="ssird" num="p" person="3" wd="ssirdn"/>
```

Figure 2. An annotation example

We have used XSLT to generate output files which allow validation of the annotated corpora. Annotation speed is between 80 and 120 tokens/hour. Randomly chosen texts were revised by three other linguists. On the basis of the revised texts inter-annotator agreement is 94.98%.Common remarks were generalized to the whole corpora in the second validation by a different annotator.

The main aim of this corpus is to learn an automatic POS tagger based on Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) because they have been proved to give good results for sequence classification (Kudo and Matsumoto, 2000, Lafferty et al. 2001). We are using freely available tools like Yamcha and CRF++ toolkits[6]. First results are very promising with more than 88% of accuracy.

5. Conclusions and future works

In this paper, after a brief description of the morpho-syntactic features of the Amazighe language and corpus encoding, we have addressed the basic principles we followed for tagging Amazighe written corpora, containing 20,667 tokens, with AnCoraPipe: the tagset used, the transliteration and the annotation tool. We plan to make available soon for research purposes the final version of the corpus.

Appendix A shows the result of applying the tagset to a sample of real Amazighe text, which proves that the defined tagset is sufficient in describing Amazighe with morpho-syntactic information.

We are planning to approach base phrase chunking by hand labeling the already annotated corpus with morphology information, afterwards to achieve an automatic base phrase chunker.

Acknowledgements

We would like to thank Kamal Ouqua, Mustapha Sghir, El hossaine El gholb and Mariam Aït Ouhssain for their precious help in the annotation task. Our thanks are also due to professors Abdallah Boumalk, Rachid Laabdallaoui and Hamid Souifi for having accepted to revise some randomly chosen annotated cases we have handed to them, and all the CAL researchers for their explanations and valuable assistance. The work of the third author was funded by the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

---

[6] Freely downloadable from http://chasen.org/~taku/software/YamCha and http://crfpp.sourceforge.net/

References

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E., Souifi, H. (2006a). Initiation à la langue Amazighe. Publications de l'IRCAM.

Ameur, M., Bouhjar, A., Boukhris, F. Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E. (2006b). Graphie et orthographe de l'Amazighe. Publications de l'IRCAM.

Andries, P. (2008). La police open type Hapax berbère. In proceedings of the workshop : la typographie entre les domaines de l'art et l'informatique, pp. 183—196.

Bertran, M., Borrega, O., Recasens, M., Soriano, B. (2008). AnCoraPipe: A tool for multilevel annotation. Procesamiento del lenguaje Natural, n° 41, Madrid, Spain.

Boukhris, F. Boumalk, A. El Moujahid, E., Souifi, H. (2008). La nouvelle grammaire de l'Amazighe. Publications de l'IRCAM.

Boumalk, A., Naït-Zerrad, K. (2009). Amawal n tjrrumt -Vocabulaire grammatical. Publications de l'IRCAM.

Civit, M. & M.A. Martí (2004). Building Cast3LB: a Spanish Treebank. In Research on Language & Computation (2004) 2, pp. 549-574. Springer, Science & Business Media. Germany.

Cohen, D. (2007). Chamito-sémitiques (langues). In Encyclopædia Universalis.

Kudo, T., Yuji Matsumoto, Y. (2000). Use of Support Vector Learning for Chunk Identification.

Lafferty, J. McCallum, A. Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In proceedings of ICML-01, pp. 282-289

Outahajala M., Zenkouar L., Rosso P., Martí A. (2010). Tagging Amazighe with AncoraPipe. In Proceedings of: Workshop on LR & HLT for Semitic Languages, 7th International Conference on Language Resources and Evaluation, LREC-2010, Malta, May 17-23, pp. 52-56.

Outahajala, M., Zenkouar, L. (2008). La norme du tri, du clavier et Unicode. In Proceedings of the workshop : la typographie entre les domaines de l'art et l'informatique, pp. 223-238.

Zenkouar L., (2008). Normes des technologies de l'information pour l'ancrage de l'écriture amazighe. Revue Etudes et Documents Berbères n°27, pp. 159-172.

Appendix A.

This paragraph is an extract from a text about "tamGra" [weddings],one of the collected texts described in the subsection 3.2. It shows the result of applying the tagset to a sample of real Amazighe text.

tlla_*ili-verb-perfective-f-s-3* **tmGra**_ *tamGra- noun-common-f-s-construct* **dar**_ *dar-prep* **wadjarn**_ *adjar-noun-common-m-p-construct* **nnG**_*nnG-det-possessive-c-p-1* **.**_*punct-period*
ira_*iri-verb-perfective-m-s-3* **urba**_ *arba-noun-commun-m-s-construct* **nnsn**_*nnsn-det-possessive-m-p-3* **ad**_ *ad-particle-preverbal* **itahl**_ *tahl- verb-aorist-m-s-3* **.**_*.-punct-period* **ar**_ *ar-particle- preverbal* **as**_ *prep-pronounpGen:c-pronounNum:s-3* **ttHyyaln**_ *Hyyl-verb-imperfective-m-p-3* **i**_*i-prep* **tmGra**_ *tamGra-noun-commun-f-s-construct* **ann**_*ann-det-demonstrative-distance* **sg**_*sg- prep* **usgg°as**_*asgg°as-noun-common-m-s-construct* **lli**_ *lli-pron-relative* **izrin**_*zri- verb-participle- perfective* **.**_*.- punct-period*
sGan_*sG-verb-perfective-m-p-3* **kigan**_*kigan-det-quantity* **n**_*n-prep*
ifckan_*afcku- noun-common-m-p -construct* **,**_ *punct-comma* **Grn**_ *Gr- verb-aorist-m-p-3* **i**_ *prep* **kigan**_ *det-quantity* **n**_*n-prep* **mddn**_ *middn- noun-common-m-p -construct* **,**_*,-punct-comma* **uggar**_*uggar-det-quantity* **n**_ *n-prep* **snat**_*sin-noun-numeral-f-p -construct* **tmaD**_*timiDi- noun-numeral-f-p -construct* **.**_*.-punct-period*

Construction et exploitation de corpus audio à l'aide du logiciel ITE

Kamal Naït-Zerrad

Inalco, Lacnad-Centre de recherche berbère (Paris)
knaitzerrad@inalco.fr

1. Corpus sonores et logiciels de saisie

Les enquêtes de terrain donnent lieu au recueil de corpus oraux qui demandent à être transcrits. L'étendue des transcriptions, traductions et annotations dépendent de l'objectif désiré. Dans notre pratique, nos besoins les plus courants sont une transcription phonétique –plus ou moins fidèle -, une transcription phonologique, une glose morphosémantique et une traduction libre. Il s'agit en effet de pouvoir travailler correctement sur ces données après l'enquête.

Il existe aujourd'hui des logiciels plus ou moins sophistiqués pour réaliser ces tâches. Ils ont chacun leur utilité selon que l'on s'intéresse plus à la phonétique, à la morphosyntaxe, etc. Citons par exemple *Praat* (www.praat.org), qui est outil pour analyser le signal acoustique ; *Transcriber* (http://trans.sourceforge.net), qui permet en particulier de transcrire des dialogues ; *Elan* (www.lat-mpi.eu/tools/elan/), logiciel très complet, qui permet l'intégration audio et vidéo et plusieurs niveaux d'analyse et *ITE*, dont il sera question ici.

2. ITE : Interlinear Text Editor

Interlinear Text Editor (ITE)[1] est un logiciel qui permet aux de saisir un corpus oral sur au moins deux niveaux : la transcription (phonétique ou phonologique) et sa glose interlinéaire. La glose morphosémantique peut être aussi détaillée que voulu. Les annotations portent sur quatre niveaux: le texte, la phrase, le mot et le morphème. Les niveaux mot et morphème présentent de manière alignée les

[1] Le logiciel ITE est librement disponible sur le site de son auteur Michel Jacobson (http://michel.jacobson.free.fr).

contenus de la transcription et de la glose. A mesure que la glose est introduite, elle est enregistrée avec la transcription correspondante sous forme de lexique. Cela permet de faciliter la saisie car dès que le programme rencontre une transcription déjà traitée, il propose la glose enregistrée. On peut donc la prendre directement ou bien en proposer une autre. Les données sont structurées dans le format XML.

ITE possède bien entendu des outils pour réaliser des concordances et des lexiques. Il permet également de faire des recherches très précises sur la transcription ou la glose. L'interrogation s'effectue à l'aide d'expressions régulières et du langage XPath sur les différents niveaux de saisie (mot, morphème, glose).

Ajoutons que le programme est très souple, on peut l'utiliser avec des DTD propres.

3. Le corpus

Il s'agit d'un extrait de corpus recueilli dans le cadre d'une thèse de doctorat (Takhedmit-Sadoudi, 2006) qui sera utilisé ici pour illustrer les possibilités d'exploitation par ITE. Il s'agit d'un entretien entre l'enquêtrice et une enquêtée en Kabylie sur des thèmes de société.

4. Exploitation du corpus

Nous allons donner des exemples montrant l'utilisation du corpus pour différentes recherches : simple, de concordances sur un élément précis, etc. Mais d'abord, voyons comment se présente les transcriptions.

4. 1. Les fenêtres de ITE

La fenêtre principale du logiciel est en fait celle de la saisie (figure 1). On voit sur cette capture d'écran un énoncé représentée par deux lignes pour le niveau « morphème » : ici la première est basée sur une transcription usuelle d'inspiration phonologique avec un découpage de mot le plus fin possible (pour un verbe par exemple : radical et indice(s) de personne). La seconde ligne constitue la glose morphosémantique la plus précise possible. Par exemple, pour un radical verbal, on peut indiquer si les thèmes d'aoriste et de prétérit sont identiques, si les deux thèmes de prétérit (positif et négatif) sont confondus, etc. Pour un nominal, on peut par exemple indiquer si les deux états (libre et annexé) sont identiques. Ces détails permettent d'effectuer des recherches les plus fines possibles.

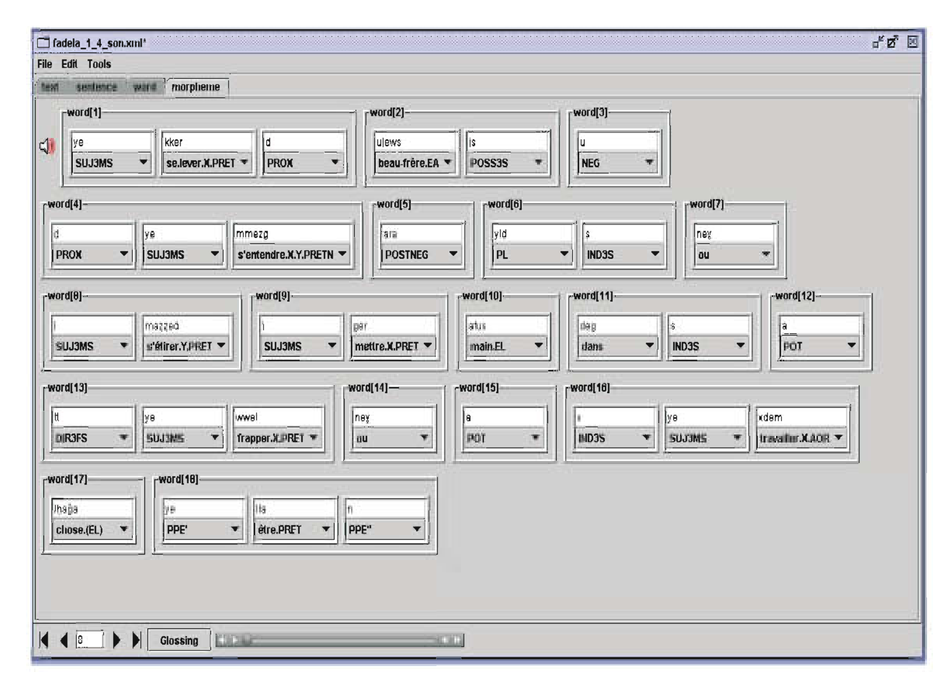


Figure 2: fenêtre principale du logiciel ITE

Pour le niveau « glose », il faut donc établir un système d'abréviations plus ou moins important selon les résultats recherchés. On peut citer encore quelques éléments qui ont une abréviation propre :

- les nominaux : l'état libre et l'état d'annexion ainsi que les cas où les lexèmes ne connaissent pas l'opposition morphologique ;

- les différents affixes et clitiques sont marqués différemment (possessifs, série directe, série indirecte, etc.)

- les verbes : On indique les verbes de qualité, les morphèmes discontinus (indices de personne, indice de participe, négation, etc.)

- et bien entendu, tous les autres éléments grammaticaux : particules, préfixes de dérivation, déictiques, etc.

Il est également possible, à partir du corpus transcrit et pré-segmenté avec un logiciel de traitement de texte quelconque de l'intégrer directement dans ITE à l'aide d'une feuille de style adéquate. Il ne reste donc plus qu'à saisir les gloses. La phrase prise comme exemple ci-dessus dans les figures 1 et 2 sera ainsi décomposée :

ye-kker-d ulews=is u d-ye-mmezg ara yid=s neɣ i-maẓẓed i-ger afus deg=s a tt=ye-wwet neɣ a s=ye-xdem lḫaǧa ye-lla-n…

Le tiret simple « - » indique un indice de personne ou de participe lié à un verbe et le signe « = » indique un affixe lié à un nom, un verbe ou une préposition.

La figure 2 montre le même énoncé que l'on peut visualiser sur le niveau « phrase ». On peut également visualiser le corpus entier sur le niveau « texte » et les mots sur le niveau « mot ».

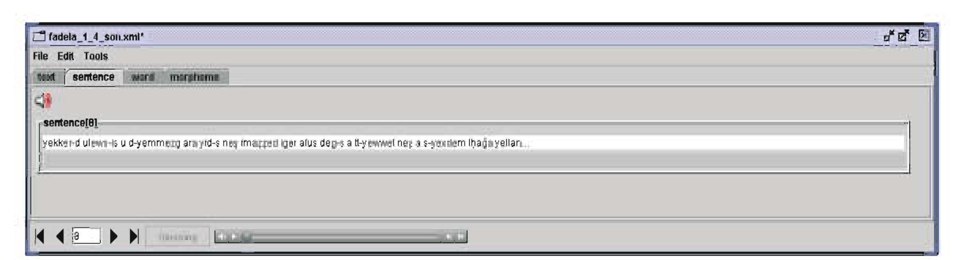


Figure 3 : fenêtre montrant la transcription d'une phrase

4.2. Concordances

Pour la linguistique, la possibilité de visualiser et de traiter des concordances précises permet d'étudier l'élément recherché dans son contexte et son cotexte. Par exemple, la figure 3 montre la recherche de la particule prédicative *d* « c'est, ce sont ». On sait qu'en berbère il existe au moins 3 homonymes (homophones et/ou homographes) pour cette particule : la particule prédicative, la particule d'orientation et la préposition. Si la glose a été bien faite, elle permet justement de différencier entre ces trois éléments. Ici, la particule prédicative est désignée par l'abréviation « PP », ce qui permet d'ajouter une expression XPath qui va, parmi les différents « d », rechercher uniquement ceux qui correspondent à la glose « PP ». Il faut également préciser que la recherche se fait au niveau morphématique et on peut ajuster le nombre de mots à droite et à gauche du mot recherché. L'expression XPath peut concerner la glose ou la transcription et peut être aussi précise qu'on le désire en employant toutes les possibilités de ce langage.

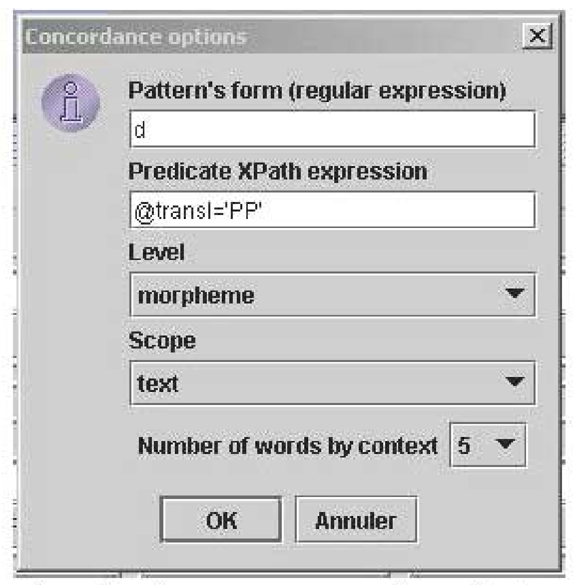
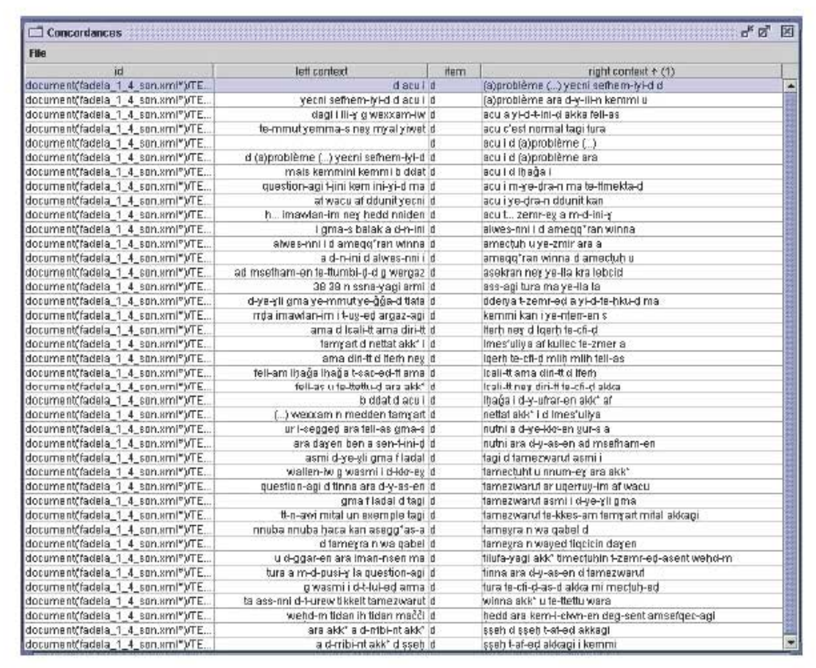


Figure 4 : options de recherche pour une concordance, ici, la particule prédicative « d »

On obtient le résultat présenté à la figure 4. Il en ressort par exemple un certain nombre de structures peu étudiées et qui ne sont pas aussi rares qu'on le pense comme les clivées à subordonnée nominale. On peut également remarquer l'utilisation de la particule préverbale « a » ou « ara » dans les clivées où le verbe de la subordonnée est à l'aoriste et dont les conditions d'apparition restent à déterminer.

4.3. Lexique

Le logiciel permet de créer un lexique de toutes les formes saisies (mots ou morphèmes) en donnant le nombre d'occurrences et bien entendu la traduction ou la glose. On peut modifier le tri : alphabétique pour les transcriptions ou les gloses, dans l'ordre croissant ou décroissant pour les occurrences, … On voit par exemple sur la figure 5 que ce sont la particule d'orientation et la particule prédicative (toutes deux « d ») qui sont les plus fréquentes.



Concordances

File

id	left context	item	right context ↑ (1)
document(fadela_1_4_son.xml")/TE...	d acu i	d	(a)problème (...) yecni sefhem-iyi-d d
document(fadela_1_4_son.xml")/TE...	yecni sefhem-iyi-d d acu i	d	(a)problème ara d-y-ili-n kemmi u
document(fadela_1_4_son.xml")/TE...	dagi i lli-ɣ g wexxam-iw	d	acu a yi-d-t-ini-d akka fell-as
document(fadela_1_4_son.xml")/TE...	te-mmut yemma-s neɣ myal ɣiwet	d	acu c'est normal tagi tura
document(fadela_1_4_son.xml")/TE...		d	acu i d (a)problème (...)
document(fadela_1_4_son.xml")/TE...	d (a)problème (...) yecni sefhem-iyi-d	d	acu i d (a)problème ara
document(fadela_1_4_son.xml")/TE...	mais kemmini kemmi i b ddat	d	acu i d lḥağa i
document(fadela_1_4_son.xml")/TE...	question-agi t-ini kem ini-yi-d ma	d	acu i m-ɣe-ḍra-n ma te-timekta-d
document(fadela_1_4_son.xml")/TE...	af wacu af ddunit yecni	d	acu i ye-ḍra-n ddunit kan
document(fadela_1_4_son.xml")/TE...	ḥ... imawlan-im neɣ hedd nniden	d	acu t... zemr-eɣ a m-d-ini-ɣ
document(fadela_1_4_son.xml")/TE...	i gma-s balak a d-n-ini	d	alwes-nni i d ameqq'ran winna
document(fadela_1_4_son.xml")/TE...	alwes-nni i d ameqq'ran winna	d	amecṭuḥ u ye-zmir ara a
document(fadela_1_4_son.xml")/TE...	a d-n-ini d alwes-nni i	d	ameqq'ran winna d amecṭuḥ u
document(fadela_1_4_son.xml")/TE...	ad msefham-en te-ttumbi-ḍ-d g wergaz	d	asekran neɣ ye-lla kra lebcid
document(fadela_1_4_son.xml")/TE...	38 39 n ssna-yagi armi	d	ass-agi tura ma ye-lla la
document(fadela_1_4_son.xml")/TE...	d-ye-ɣli gma ye-mmut ye-ǧǧa-d tlata	d	dderya t-zemr-eḍ a yi-d-te-ḥku-d ma
document(fadela_1_4_son.xml")/TE...	mḍa imawlan-im i t-uɣ-eḍ argaz-agi	d	kemmi kan i ye-nṭerr-en s
document(fadela_1_4_son.xml")/TE...	ama d lcali-tt ama diri-tt	d	lferḥ neɣ d lqerḥ te-cfi-ḍ
document(fadela_1_4_son.xml")/TE...	tamɣart d nettat akk' i	d	lmes'uliya af kullec te-zmer a
document(fadela_1_4_son.xml")/TE...	ama diri-tt d lferḥ neɣ	d	lqerḥ te-cfi-ḍ mlih mlih fell-as
document(fadela_1_4_son.xml")/TE...	fell-am lḥağa lḥağa t-sac-ed-tt ama	d	lcali-tt ama diri-tt d lferḥ
document(fadela_1_4_son.xml")/TE...	fell-as u te-ttettu-ḍ ara akk'	d	lcali-tt neɣ diri-tt te-cfi-ḍ akka
document(fadela_1_4_son.xml")/TE...	b ddat d acu i	d	lḥağa i d-y-ufrar-en akk' af
document(fadela_1_4_son.xml")/TE...	(...) wexxam n medden tamɣart	d	nettat akk' i d lmes'uliya
document(fadela_1_4_son.xml")/TE...	ur i-segged ara fell-as gma-s	d	nutni a d-ye-kks-en ɣur-s a
document(fadela_1_4_son.xml")/TE...	ara dayen ben a sen-t-ini-ḍ	d	nutni ara d-y-as-en ad msefham-en
document(fadela_1_4_son.xml")/TE...	asmi d-ye-ɣli gma f ladal	d	tagi d tamezwarut asmi i
document(fadela_1_4_son.xml")/TE...	wallen-iw g wasmi i d-kks-eɣ	d	tamecṭuḥt u nnum-eɣ ara akk'
document(fadela_1_4_son.xml")/TE...	question-agi d tinna ara d-y-as-en	d	tamezwarut ar ugerruy-im af wacu
document(fadela_1_4_son.xml")/TE...	gma f ladal d tagi	d	tamezwarut asmi i d-ye-ɣli gma
document(fadela_1_4_son.xml")/TE...	tt-n-awi mital un exemple tagi	d	tamezwarut te-kkes-am tamɣart mital akkagi
document(fadela_1_4_son.xml")/TE...	nnuba nnuba ḥaca kan asegg'as-a	d	tameɣra n wa qabel d
document(fadela_1_4_son.xml")/TE...	d tameɣra n wa qabel	d	tameɣra n wayed tiqcicin dayen
document(fadela_1_4_son.xml")/TE...	u d-ggar-en ara iman-nsen ma	d	tilufa-yagi akk' timecṭuḥin t-zemr-eḍ-asent weḥd-m
document(fadela_1_4_son.xml")/TE...	tura a m-d-pusi-ɣ la question-agi	d	tinna ara d-y-as-en d tamezwarut
document(fadela_1_4_son.xml")/TE...	g wasmi i d-t-lul-eḍ arma	d	tura te-cfi-ḍ-as-d akka mi mecṭuḥ-eḍ
document(fadela_1_4_son.xml")/TE...	ta ass-nni d-t-urew tikkelt tamezwarut	d	winna akk' u te-ttettu wara
document(fadela_1_4_son.xml")/TE...	weḥd-m tidan ih tidan maččči	d	hedd ara kem-i-ckwn-en deg-sent amsefqec-agi
document(fadela_1_4_son.xml")/TE...	ara akk' a d-mbi-nt akk'	d	ṣṣeḥ d ṣṣeḥ t-af-ed akkagi
document(fadela_1_4_son.xml")/TE...	a d-mbi-nt akk' d ṣṣeḥ	d	ṣṣeḥ t-af-eḍ akkagi i kemmi

Figure 5 : résultat de la recherche de concordances

morpheme lexicon

File

transcription	gloss	occurrences ↓ (1)
d	PROX	41
d	PP	40
i	RELI	32
a	POT	32
eḍ	SUJ2S"	28
ye	SUJ3MS	28
t	SUJ2S'	28
u	NEG	20
te	SUJ2S'	18
ara	POSTNEG	17
ḍ	SUJ2S"	17
neɣ	ou	17
te	SUJ3FS	16
ini	dire.AOR	14
g	dans	13
ɣ	SUJ1S	12
kem	DIR2FS	12
akk'	tout	12
ara	RELA	10
fell	sur	10
zemr	pouvoir.X.PRET	10
ma	si	9
lla	être.PRET	9
im	POSS2FS	9

Figure 6 : Extrait du lexique engendré par ITE à partir du corpus transcrit, ici par ordre décroissant du nombre d'occurrences.

4.4. Intégration multimédia

L'enregistrement sonore peut être associé au texte à l'aide de l'éditeur de texte XML et de son SoundIndex (SI.TCL) réalisé en Tcl/Tk[2]. Il permet d'insérer des balises audio à un niveau quelconque du document XML.

Le document final (texte + son) peut alors être lu par ITE phrase par phrase, ou même à un niveau inférieur selon l'objectif désiré : les balises peuvent être intégrées au niveau de la phrase ou d'un mot. Bien entendu, on peut lire et écouter tout le corpus dans l'onglet « texte ».

Étant donné la flexibilité de ITE, on peut également intégrer un enregistrement vidéo.

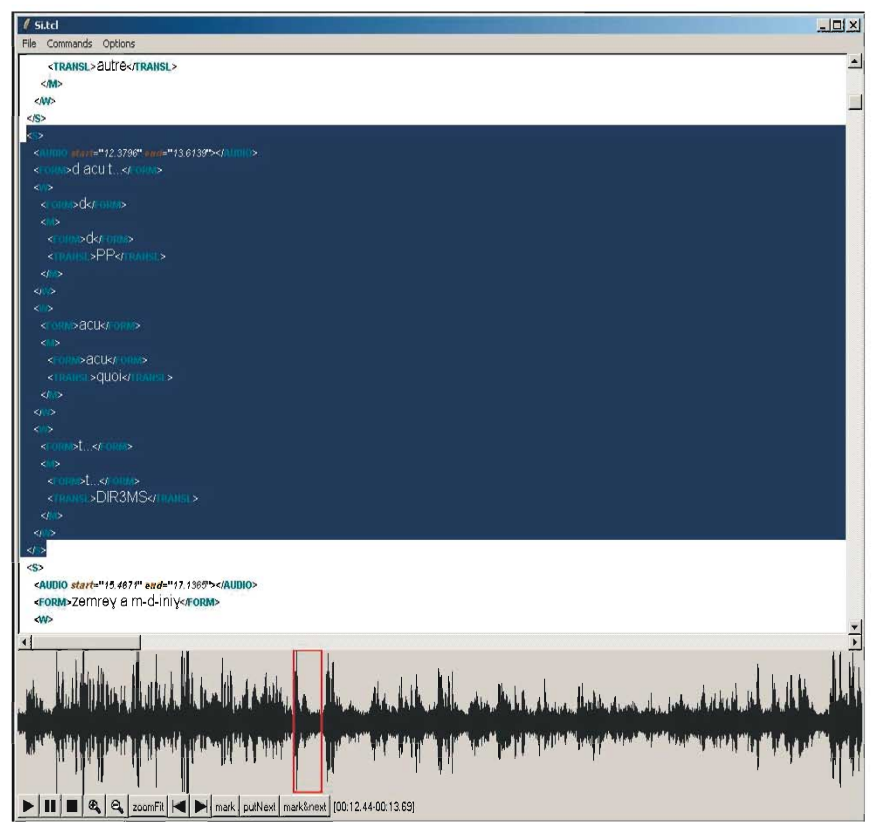


Figure 7 : fenêtre de l'éditeur SiTcl

[2] Le logiciel SounIndex est librement disponible sur le site de son auteur Michel Jacobson (http://michel.jacobson.free.fr)

Dans l'exemple présenté en figure 6, nous avons découpé l'enregistrement par « phrases », délimitées par la balise <S> et </S>. Le logiciel permet donc de marquer chaque phrase à partir de la séquence audio correspondante et attribue automatiquement la balise <audio> indiquant le début et la fin de la séquence. Les figures 1 et 2 montrent comment le son est matérialisé dans la fenêtre principale.

4.5. Autres possibilités

On peut créer des feuilles de styles (xslt) et les appliquer pour obtenir différentes présentations, transformations ou découpage du texte. Par exemple, on peut obtenir une présentation juxtalinéaire de la transcription et de la glose, directement à partir de ITE (il s'agit de la phrase prise comme exemple dans la fenêtre ITE, figure 1) :

ye-kker-d	**ulews-is**	**u**	**d-ye-mmezg**
SUJ3MS-se.lever.X.PRET-PROX	beau-frère.EA-POSS3S	NEG	PROX-SUJ3MS-s'entendre.X.Y.PRETN

ara	**yid-s**	**neɣ**	**i-maẓẓed**	**i-ger**	**afus**
POSTNEG	PL-IND3S	ou	SUJ3MS-s'étirer.Y.PRET	SUJ3MS-mettre.X.PRET	main.EL

deg-s	**a**
dans-IND3S	POT

tt-ye-wwet	**neɣ**	**a**	**s-ye-xdem**	**lḥağa**
DIR3FS-SUJ3MS-frapper.X.PRET	ou	POT	IND3S-SUJ3MS-travailler.X.AOR	chose.(EL)

ye-lla-n
PPE'-être.PRET-PPE"

5. Conclusion

Comparé à un logiciel sophistiqué comme *ELAN*, *ITE* possède certaines limites comme le nombre de niveaux d'analyse. Cependant, il présente l'avantage de la simplicité d'utilisation avec des possibilités d'exploitation et d'analyse étendues. En effet, comme nous l'avons indiqué plus haut, la glose morphosémantique peut être très fine, elle peut associer un élément de transcription à une abréviation qui permet de le retrouver dans son contexte sans ambigüité possible avec un homonyme. La précision des gloses permet également de faire diverses statistiques sur les unités lexicales et grammaticales.

Un autre avantage du logiciel tient dans l'écriture des gloses. A partir d'un corpus assez important, les gloses se font pratiquement automatiquement puisque ITE les enregistre au fur et à mesure de leur saisie en les associant à la transcription correspondante. Dès que ITE rencontre une transcription avec une glose connue, il la propose : soit elle est acceptée par l'utilisateur soit il fournit une autre glose.

Ce logiciel nous semble donc tout à fait adéquat comme aide à l'analyse linguistique des corpus oraux.

Références

Takhedmit-Sadoudi H. (2006). *Récits de vie et points de vue dans les dialogues avec des femmes kabyles : analyse discursive*. Thèse de doctorat, Paris 5, Dir. F. François, K. Naït-Zerrad.

Jacobson M. (2004). Corpus oraux glosés: outils logiciels d'aide à l'analyse. In Purnelle G., Fairon C. & Dister A. (éds), *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles, Vol. II*, pp. 625-632.

Outahajala M., Zenkouar L., Rosso P., Martí A. (2010).Tagging Amazigh with AncoraPipe. In: Proc. Workshop on LR & HLT for Semitic Languages, 7th Int. Conf. on Language Resources and Evaluation, LREC-2010, Malta, May 17-23, pp. 52-56.

Tesniere,L. (1959). *Eléments de syntaxe structurale*- édition (1966) (1969) (1988). Klincksieck- Paris.

5. خلاصة

انطلاقا من المقطع المدروس يمكن القول إن أية حوسبة تروم معالجة السلوك التركيبي (le comportement syntaxiques) لمقولات المتون اللغوية ، وتسعى أن تحقق للمستعمل مبدأ البساطة و الدقة، لا بد لها من التركيز على:

- تحديد المعلومات الصرفية التي تحدد البنية الصرفية لكل المقولات التي تتألف منها الجمل؛
- تحديد التقطيع التركيبي الذي يحدد المكونات الكبرى للجمل؛
- تحديد المكملات التي تمكن من تحديد الاشتغال التركيبي الدلالي لمركبات الجمل.

المصادر والمراجع المعتمدة

أقا، كمال.(2009). *محلاتية الفعل وضوابط تأليف المعاجم التعليمية الثنائية اللغة- مشروع معجم عربي/ أمازيغي نموذجا*- رسالة لنيل شهادة الدكتوراه في اللسانيات. كلية الآداب. مكناس.

فنان، أمينة .(1997). المكونات التكميلية للجملة الفعلية، التوسعات. ندوة مكانة الأنحاء التقليدية في اللسانيات الحديثة. سلسلة الندوات. كلية الآداب. مكناس

مسلك، عبد الرزاق .(1995). نظرية محلاتية الفعل و أهميتها في تدريس اللغات الأجنبية. مجلة كلية الآداب. ظهر المهراز، فاس

Allerton,D. (1996). *Valency and valency Grammar, In Concise encyclopedia of syntactic theories-University of Edingburg*. Cambridge university press.

Boukhris, F. Boumalk, A. El moujahid, E., Souifi, H. (2008). *La nouvelle grammaire de l'Amazighe*. Publications de l'IRCAM- El Maârif Al jadida- Rabat

Engel.U. (1977). *Syntax der deutschen Gegenwatssprache.* Erich Shmidt Verlag-Berlin.

Kaddour,C (1987) *Le système verbal rifain: forme et sens*. SELAf, Paris.

Kaddour,C. (1990). *Transitivité et Diathèse en Tarifit : analyse de quelques relation de dépendance lexicale et syntaxique- thèse de doctorat d'état* .Université de Sorbonne nouvelle- Paris3.

<n gen="m" lem="azgn" num="s" postype="common" state="construct" wd="uzgn"/>

<s lem="n" wd="n"/>

<n gen="m" lem="ass" num="s" postype="common" state="construct" wd="wass"/>

</grup.prep>

<f punct="period" wd="."/>

الجملة الخامسة:

<u>ⵏⵎⵎⴽⵍⵓ ⵖⴰⵔ ⵓⵎⴷⴷⴰⴽⴽⵯⵍ ⵏ ⴱⴰⴱⴰ.</u>

تتألف هذه الجملة من ثلاث مركبات هي:

- مركب فعلي وهو: ⵏⵎⵎⴽⵍⵓ
- مركب حرفي وهو : ⴱⴰⴱⴰ ⵏ ⴰⵎⴷⴷⴰⴽⴽⵯⵍ ⵖⴰⵔ الذي يؤدي وظيفة المكمل المكاني

<sentence>

<grup.verb Typecompl="subjectival">

<v aspect="perfective" gen="c" lem="mmklu" num="p" person="1" wd="nmmuklu"/>

</grup.verb>

<grup.prep Typecompl="locative">

<s lem="Gar" wd="Gar"/>

<n gen="m" lem="amddakl" num="s" postype="common" state="construct" wd="umddukl"/>

<s lem="n" wd="n"/>

<n gen="m" lem="bab" num="s" person="1" possessorgen="c" possessornum="s" postype="parental" wd="baba"/>

</grup.prep>

<f punct="period" wd="."/>

</sentence>

- <u>ⵏⵟⵟⴼ ⴰⴱⵔⵉⴷ,</u>

```
<grup.verb Typecompl=”subjectival”>
    <v aspect=”imperfective” gen=”c” lem=”TTf” num=”p” person=”1”
voice=”active” wd=”nTTf”/>
</grup.verb>
<grup.nom Typecompl=”locative” >
    <n Typecompl=”locative” gen=”m” lem=”abrid” num=”s”
    postype=”common” state=”free” wd=”abrid”/>
</grup.nom>
<f punct=”comma” wd=”,”/>
```

الجملة الرابعة:

- <u>ⵏⵉⵡⴹ ⵖⵔ ⵣⴰⴳⵓⵔⴰ ⵖⵔ ⵓⵣⴳⵏ ⵏ ⵡⴰⵙⵙ,</u>

تتألف هذه الجملة من ثلاث مركبات هي:

- مركب فعلي وهو: ⵏⵉⵡⴹ
- مركب حرفي وهو:ⵖⵔ ⵣⴰⴳⵓⵔⴰ الذي يؤدي وظيفة المكمل المكاني
- مركب حرفي وهو: ⵖⵔ ⵓⵣⴳⵏ ⵏ ⵡⴰⵙ الذي يؤدي وظيفة المكمل الزماني

```
<grup.verb Typecompl=”subjectival”>
    <v aspect=”perfective” gen=”c” lem=”awD” num=”p” person=”1”
    voice=”active” wd=”niwD”/>
</grup.verb>
<grup.prep Typecompl=”locative”>
    <s lem=”Gr” wd=”Gr”/>
    <n gen=”f” lem=”zagura” num=”s” postype=”proper” state=”free”
wd=”zagura”/>
</grup.prep>
<grup.prep Typecompl=”locative”>
    <s lem=”Gr” wd=”Gr”/>
```

```
<s lem="xf" wd="xf"/>
<n Typecompl="dative" gen="m" lem="amuddu" num="s" postype="common" state="construct" wd="umuddu"/>
<d gen="c" lem="nns" num="s" person="3" postype="possessive" wd="nns"/>
</grup.prep>
<f punct="colon" wd=":"/>
</sentence>
```

الجملة الثانية:

<u>ⵏⴽⴽⵔ ⵣⵉⴽⴽ,</u>

تتألف هذه الجملة من مركبين هما:

مركب فعلي وهو: ⵏⴽⴽⵔ

مركب ظرفي وهو: ⵣⵉⴽⴽ الذي يؤدي وظيفة المكمل الحر.

وتتمثل معلوماتيا على هذا النحو:

```
<sentence>
<grup.verb Typecompl="subjectival">
  <v aspect="perfective" gen="c" lem="kkr" num="p" person="1" voice="active" wd="nkkar"/>
</grup.verb>
<grup.adv Typecompl="free">
  <ad lem="zikk" postype="time" wd="zikk"/>
</grup.adv>
<f punct="comma" wd=","/>
```

الجملة الثالثة:

تتألف هذه الجملة من مركبين هما:

مركب فعلي هو: ⵏⴹⴹⵃ

مركب اسمي هو: ⴰⵙⵔⵉⴷ الذي يؤدي وظيفة المكمل المكاني.

وفيما يلي التمثيل المعلوماتي[6] (la représentation informatique) لمحلاتية جمل المقطع المدروس([7]).

الجملة الأولى:

<u>ⵀⴰ ⵜⵉⴼⴰⵡⵜ ⵜⵜⴰⵍⵙ ⵅⴼ ⵓⵎⵓⴷⴷⵓ ⵏⵏⵙ.</u>

تتألف هذه الجملة من ثلاث مركبات هي:

المركب الاسمي: وهو ⵀⴰ ⵜⵉⴼⴰⵡⵜ ويؤدي في هذه الجملة وظيفة المكمل الفاعلي.

المركب الفعلي: وهو ⵜⵜⴰⵍⵙ ([8])

المركب الحرفي: وهو (ⵅⴼ ⵓⵎⵓⴷⴷⵓ ⵏⵏⵙ) ويؤدي في هذه الجملة وظيفة المكمل الإضافي لفعل القول (ⴰⵍⵙ).

وتتمثل معلوماتيا على هذا النحو:

```
<sentence>
    <grup.nom Typecompl="subjectival">
        <d lem="ha" postype="demonstrative" wd="ha"/>
        <n Typecompl="subjectival" gen="f" lem="tifawt" num="s" postype="proper" state="free" wd="tifawt"/>
    </grup.nom>
    <grup.verb Typecompl="subjectival">
        <v aspect="imperfective" gen="f" lem="als" num="s" person="3" voice="active" wd="ttals"/>
    </grup.verb>
    <grup.prep Typecompl="dative">
```

[6] نشير هنا إلى أن البرنامج المعتمد قد عدل وفق ما يقتضيه الوسم المحلاتي للأمازيغية بمساعدة محمد أتهجلا، باحث بالمعهد الملكي للثقافة الأمازيغية.

[7]-يعتمد البرنامج المعلوماتي Ancorapipe لغة وصف البيانات (XML)

[8] - هذا المركب يتألف من فعل وعائد ضميري عن مكمل فاعلي، و سنعالج معلوماتيا عائديات المقطع المدروس في بحث لاحق.

Grup.nom	Nominal group
Grup.prepo	Prepositional group
Grup.verb	Verbal group

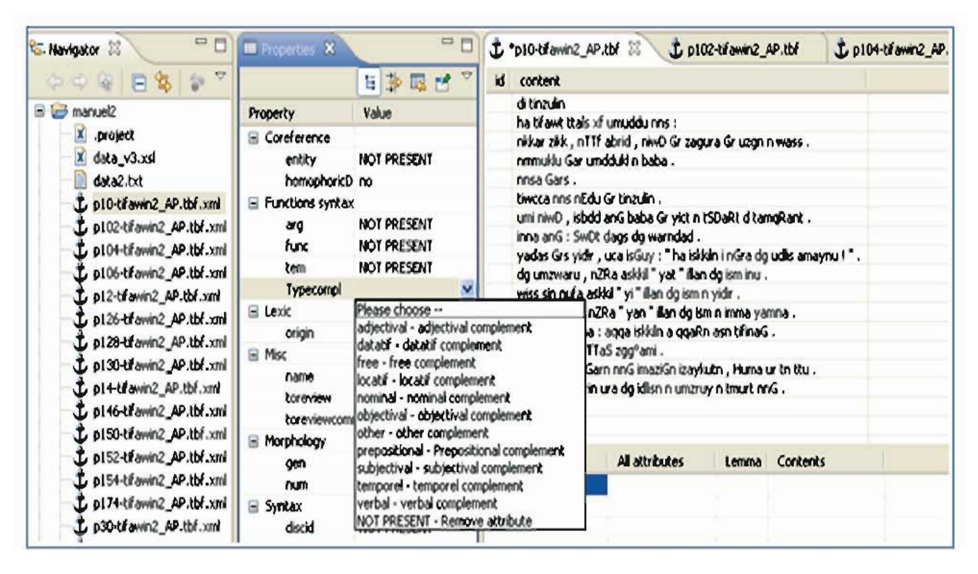
جدول توضيحي1. المختصرات والمركبات التي تدل عليها في البرنامج.

2 - تحديد صنف المكمل (type complement) المميز لكل مركب حسب الجملة المعالجة، على أن يتم تعزيز هذه المكملات بعائديتها واستبداليتها ونوعها في بحث لاحق.

Etiquette utilisée	مدلولها
adjectival	المكمل الوصفي
dative	المكمل الإضافي
free	المكمل الحر
locative	المكمل المكاني
nominal	المكمل الاسمي
objectival	المكمل المفعولي
prepositional	المكمل الحرفي
subjectival	المكمل الفاعلي
temporal	المكمل الزماني
verbal	المكمل الجملي

جدول توضيحي 2. المصطلحات المحلاتية الواردة في البرنامج.

المدرسي (ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ) للسنة الثانية و الموسوم صرافيا (Etiquetage morphosyntaxique) بواسطة البرنامج المعلوماتي (Ancorapipe)([3]).



وثيقة1. مقطع لنص موسوم محلاتيا ([4])

واعتمادا على التقطيع التركيبي للمقطع المدروس (la ségmentation syntaxique) حددنا لكل جملة:

1 - المركبات التي تتألف منها، والتي تنحصر عادة في: المركب الفعلي (verbal group)، المركب الاسمي(nominal group)، المركب الوصفي(adjectival group)، المركب الحرفي (prepositional group)، المركب الظرفي(adverbial group)([5]).

Abréviation	Signification
Grup.adverb	Adverbial group

3- لمزيد من المعلومات حول الوسم الصرافي للغة الأمازيغية باستعمال هذا البرنامج يراجع أتهجلا وآخرون (2010)

4- النص مأخوذ من الكتاب المدرسي (ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ) السنة الثانية ص 10.

5- لمزيد من الاطلاع على بنية هذه المركبات في اللغة الأمازيغية والوحدات الصرفية التي تشغلها تركيبيا يراجع بوخريص وآخرون (2008)

- ⵉⵙⵙⵉⵡⴹ ⵏ ⵓⵎⵖⴰⵔ ⵉ ⵎⵉⴷⵏ ⵉⵙ ⵉⵔⴰ ⴰⴷ ⴷ ⵉⴷⴷⵓ ⵓⴳⵍⵍⵉⴷ (أبلغ الشيخ الناس بأن الملك سيأتي).

- ⵖⴰⵍⵏ ⵉⵏⵍⵎⴰⴷⵏ ⵉⵙ ⵓⵔ ⴷ ⵉⴷⴷⵉ ⵓⵙⵍⵎⴰⴷ (ظن التلاميذ أن الأستاذ لن يأتي)

3. الأدوار الوظيفية و الأبعاد التعليمية للحوسبة المحلاتية للأفعال في الكتاب المدرسي للأمازيغية

تعتبر محلاتية الفعل من بين الإطارات اللسانية التي يمكن استثمار نتائج تحليلها في تعليم وتعلم اللغة وتبرز أدوارها التعليمية انطلاقا من إيلائها الأهمية لـ:

- السياق في تحديد مكملات أفعال لغة من اللغات الطبيعية، مما يسمح بتجاوز الطرائق التعليمية القائمة على تدريس وحدات المعجم باعتبارها وحدات مستقلة؛

- إبراز التعالقات المعجمية بين الفعل ومكملاته، مما يمكن المستعمل من إدراك أن الوحدات المعجمية مبهمة المعنى ما لم تورد تركيبيا مع وحدات أخرى؛

- الفروق بين المكملات الضرورية والاختيارية والحرة للأفعال حسب سياقات ورودها، مما يمكن متعلم اللغة الأمازيغية من التمييز بين الأفعال التي تنتمي إلى الاشتراك الدلالي (polysémie) و التجانس (homonymie) والأفعال التي تحمل مفعول المعنى (effet de sens) (قدور، 1990: 186-195).

- تقليص مجموعة من الأبواب النحوية وإعادة تبويبها.

- التمييز بين الظروف المرتبطة بالأفعال (مكمل مكاني) (مكمل زماني) وغير المرتبطة بالأفعال ضرورة.

4. التطبيقات الحاسوبية للمحلاتية انطلاقا من الكتاب المدرسي للأمازيغية " ⵜⵉⵣⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ "

يتيح البناء العام لمحلاتية الفعل عند إنغل إمكانية حوسبة النظام اللغوي لأية لغة طبيعية وذلك لكونها تتوفر على قواعد صورية قادرة على معالجة سائر الجمل، هذه القواعد تتخذ شكل نحو صوري يهدف إلى إخراج كل ما يشق على نظام اللغة والاحتفاظ فقط بما هو مقبول تركيبيا و دلاليا.و انطلاقا من أهميتها هاته سنحاول أن نمهد لتطبيقها على أفعال اللغة الأمازيغية انطلاقا من نص قرائي مقتطف من الكتاب

- ⵢⵓⵍⵙ ⵓⵎⵖⴰⵔ ⵉ ⵉⵛⵉⵔⵔⴰⵏ ⵜⵉⵏⴼⵓⵙⵉⵏ ⵜⵉⵣⴰⵢⴽⵓⵜⵉⵏ (حكى الشيخ للأطفال قصصا قديمة)

المكمل المكاني(le complément locatif)، ويتعلق الأمر بالمكمل الذي يدل على مكان وقع فيه أو عليه الحدث، نحو:

- ⵉⵣⴷⵖ ⵓⵔⴳⴰⵣ ⴳ ⵜⴰⴷⴷⴰⵔⵜ ⵏⵏⵙ (سكن الرجل في منزله)

- ⵢⴰⵖⵓⵍ ⴷ ⵃⴷⴷⵓ ⴳ ⴼⵔⴰⵏⵙⴰ ⵖⵔ ⵜⵎⴰⵣⵉⵔⵜ ⵏⵏⵙ (رجع حدو من فرنسا إلى بلدته).

المكمل الزماني(le complément temporel)، ويتعلق الأمر بالمكمل الذي تنتقيه بعض الأفعال كمكمل ثان للدلالة على زمن وقع فيه واستوجبه الحدث،نحو:

- ⵉⴽⴽⴰ ⵏ ⵙⵉⴼⴰⵡ ⴰⵙⴳⴳⵯⴰⵙ ⴳ ⵉⵎⵓⴷⴷⵓⵜⵏ ⵏⵏⵙ (قضى سيفاو سنة في أسفاره)

- ⵜⵇⵇⵉⵎⴰ ⵜⴰⵖⴰⵔⵜ ⴳ ⵜⵎⴰⵣⵉⵔⵜ ⵉⵙⴳⴳⵯⴰⵙⵏ (استمر الجفاف في البلدة سنوات)

المكمل الاسمي(le complément nominal)، ويتعلق الأمر بالأسماء الجامدة التي تأتي خبرا في ما يسمى في الأنحاء القديمة بالجملة الاسمية([2]) نحو:

- ⴰⵔⴳⴰⵣ ⴷⴷⵖ ⴷ ⴰⵎⴰⵣⵉⵖ (هذا الرجل أمازيغي)

- ⵜⴰⵙⴽⵍⵓⵜ ⴷⴷⵖ ⴷ ⵜⴰⵔⵎⵎⴰⵏⵜ (هذه الشجرة رمانة)

المكمل الوصفي(le complément adjectival)، ويتعلق الأمر بالصفات الأمازيغية (أسماء الفاعلين، والمفعولين، وصيغ المبالغة) التي تأتي خبرا في ما يسمى في الأنحاء القديمة بالجملة الاسمية، نحو:

- ⵎⵓⵃⴰ ⴷ ⴰⵎⴽⵔⴰⵣ (موحى حراث)

- ⵄⵍⵉ ⴷ ⴰⵎⴰⵣⴰⵏ ⵏ ⵜⵇⴱⵉⵍⵜ (علي مبعوث القبيلة)

المكمل الجملي. ويتعلق الأمر بالجمل التي تأتي مكملا ثانيا أو ثالثا لأفعال من قبيل (ⵉⵏⵉ) (ⵙⵙⵉⵡⴹ) (ⵖⴰⵍ) كما في الأمثلة التالية:

- ⵉⵏⵏⴰ ⴰⵙ ⴱⴰⵙⵙⵓ ⵉ ⵎⵓⵃⴰ ⵀⴰⵏ ⵃⴷⴷⵓ ⵉⴷⴷⴰ ⴷ (قال باسو لموحى إن حدو قد جاء)

---

[2]- يدمج المحلاتيون الجمل الاسمية كذلك في إطار محلاتية الفعل، نظرا لأنها جمل يقودها في الأصل فعل الكون (ⴳ) أو إحدى مرادفاته، لمزيد من المعلومات، يراجع أليرتون (1996: 366-367)

- المكملات الحرة(les compléments libres)، وهي المكملات التي لا يقتضيها الفعل لا على وجه اللزوم، ولا على وجه الاختيار، بل تذكر في الجملة دون شرط (أقا،158:2009).

2.2. أصناف المكملات عند إنغل

المكمل(Complément) عند إنغل(1977) هو كل عنصر يتبع مباشرة صنفا من الأفعال على الشكل الذي يسمح بتقسيم هذا الصنف من الأفعال إلى عدد من أشباه الأصناف، نحو الأفعال التي تحتاج مكملا زمانيا، أو مكانيا، أو إضافيا. ومن خلال تطبيقه لعمليات تشخيص مكملات أفعال اللغة الألمانية حدد إنغل تسعة أصناف من مكملات الفعل تتمظهر في اللغة الأمازيغية في ما يلي:

- المكمل الفاعلي(le complément subjectival)، وهو كل اسم ذكرته بعد فعل وأسندت ذلك الفعل إلى الاسم، ويشتمل هذا الصنف من المكملات في اللغة الأمازيغية على الأسماء ومعوضاتها التي تأتي فاعلا ونائبا للفاعل، نحو:

-ⵢⴰⵔⵓ ⵓⵙⵍⵎⴰⴷ ⵜⴰⴱⵔⴰⵜ(كتب الأستاذ رسالة)

- ⵜⵜⵢⴰⵔⵓ ⵜⴱⵔⴰⵜ (كتبت الرسالة)

- المكمل المفعولي(le complément objectival)، وهو كل اسم يأخذ الوظيفة النحوية المسماة في الأنحاء القديمة المفعول المباشر، نحو:

-ⵉⵙⵖⴰ ⵓⵔⴳⴰⵣ ⵜⵉⵟⵉⵔⵉⵜ(اشترى الرجل السيارة)

- ⵉⵜⵛⴰ ⵓⵔⴱⴰ ⵉⵙⵍⵎⴰⵏ (أكل الولد السمك)

- المكمل الحرفي(le complément prépositionnel)، وهو كل اسم يأخذ الوظيفة النحوية المسماة في الأنحاء القديمة المفعول غير المباشر للأفعال من قبيل (ⵙⵉⴽⴽ) و(ⴳ) كما في المثالين التاليين:

-ⵉⵙⵉⴽⴽ ⵓⵙⵍⵎⴰⴷ ⵅⴼ ⵉⵙⵉⵔⴰⵎⵏ ⵏ ⵉⵏⵍⵎⴰⴷⵏ(صحح الأستاذ تمارين التلاميذ)

- ⵉⴳⴰ ⴰⵙ ⵎⵓⵃⴰ ⵉ ⵎⵉⵎⵎⵉⵙ ⵙⵉⴼⴰⵡ(سمى موحى ابنه سيفاو)

- المكمل الإضافي(le complément datif)، ويتعلق الأمر بالمكمل الثاني أو الثالث الذي يأتي بعد حرف الجر(ⵉ) لأفعال القول والعطاء والتأثير ، نحو:

-ⵉⴼⴰ ⵓⵀⵔⵔⵉⵎ ⵜⴰⵏⵡⵡⴰⵛⵜ ⵉ ⵜⵎⴷⴷⴰⴽⵍⵜ ⵏⵏⵙ(أعطى الشاب الوردة لحبيبته)

- ⵢⵓⴽⵔ ⵓⵎⵅⵅⴰⵔ ⵉⴷⵔⵉⵎⵏ ⵉ ⵓⵙⴱⴱⴰⴱ(سرق اللص الدراهم للتاجر)

1– ⵉⵜⵛⴰ ⵎⵓⵃⴰ ⴰⴽⵙⵓⵎ (أكل موحى اللحم)

1.1– ⵉⵜⵛⴰ ⴰⵜ ⵎⵓⵃⴰ (أكله موحى)

2– ⵉⴼⴰ ⵓⵙⵍⵎⴰⴷ ⵜⴰⵙⵎⵖⵓⵔⵜ ⵉ ⵓⵏⵍⵎⴰⴷ (أعطى الأستاذ التلميذ جائزة)

2.1– ⵉⴼⴰ ⴰⵙ ⵜⴰⵙⵎⵖⵓⵔⵜ (أعطاه جائزة)

3 - ⵜⴷⴷⴰ ⵜⵏⵍⵎⴰⴷⵜ ⵖⵔ ⵜⵉⵏⵎⵍ (ذهبت التلميذة إلى المدرسة)

3.1– ⵜⴷⴷⴰ ⵜⵏⵍⵎⴰⴷⵜ ⵣⵉⵔⵕ (ذهبت التلميذة إليها)

- الحذف (l'élimination) وهي عملية يولد من ورائها التفريق بين عناصر الجملة التي تتبع الفعل والتي لا تتبعه بحال من الأحوال، أما وظيفتها الثانية فتتجسد في أنها تساعد على معرفة العناصر الضرورية و الاختيارية، ومن ثم فإن إنغل اعتمد الحذف كرائز للتعرف على نحوية الجملة، وتصبح العناصر المحصل عليها بعد هذه العملية ضرورية للتكوين التركيبي للجملة، وكمثال على ذلك، إن المكمل المفعولي في الجملتين (4) و (5) يعتبر ضروريا بالنسبة للفعل (ⵣⵣⵏⵣ) وحذفه يولد جملة غير مقبولة تركيبيا ودلاليا، واختياريا بالنسبة للفعل (ⴷⴷⵓ) حيث إن حذفه يبقي على الجملة صحيحة تركيبيا ودلاليا.

4- ⵣⵣⵏⵣⴰⵏ ⵉⵎⵙⴷⵓⵔⵔⵓⵢⵏ ⵜⴰⵎⴰⵣⵉⵔⵜ (باع المستعمرون البلدة)

4.1- ⵣⵣⵏⵣⴰⵏ ⵉⵎⵙⴷⵓⵔⵔⵓⵢⵏ ? (باع المستعمرون)

5- ⵉⴷⴷⴰ ⵎⵓⵃⴰ ⵖⵔ ⵜⵎⴰⵣⵉⵔⵜ (ذهب موحى إلى البلدة)

5.1 ⵉⴷⴷⴰ ⵎⵓⵃⴰ (ذهب موحى)

وانطلاقا من أن هناك تركيبات نحوية ثانوية يمكن تجاوزها واستبعادها تواصليا دون أن يتأذى من ذلك القصد الإخباري. يفرق إنغل بين المكملات الإجبارية والثانوية والحرة.

- المكملات الإجبارية(les compléments obligatoires)، وهي المكملات التي يقتضيها الفعل على وجه اللزوم، فإذا ما حذفت تبقى الجملة غير صحيحة تركيبيا و دلاليا.

- المكملات الاختيارية(les compléments facultatifs)، وهي المكملات التي لا يقتضيها الفعل على وجه اللزوم، بل على وجه الاختيار، فإذا ما حذفت تبقى الجملة صحيحة تركيبيا.

ويعتبر مفهوم تنيير للمحلاتية المنطلق الأساس الذي سيطور على يد مجموعة من اللسانيين الأوروبيين والأمريكيين، ابتداء من ستينيات القرن الماضي، إذ أدمج مجموعة من الباحثين الألمان مفهوم المحلاتية ضمن نظرية أكبر هي نظرية التبعية la théorie de dépendance وبرز داخل الاتجاه الألماني نموذج مهم للساني إنغل (Engel)استطاع أن يركز على الجوانب التركيبية والدلالية للأفعال، فحسب إنغل لا تحدد المحلاتية عدد المحلات فقط، كما ادعى ذلك تنيير([1])، بل تحدد كذلك أنواعها. وبهذا يصح القول إن الفعل يعطي معلومات عن البنية الصرف- تركيبية لعناصر الجملة التي تتبع الفعل مباشرة.

1.2.أنواع المكملات عند إنغل

اعتمد إنغل على ثلاث عمليات لسانية لضبط أنواع مكملات الأفعال في اللغات الطبيعية، وهي:

-الاستبدال (la substitution) وذلك قصد معرفة عناصر الجملة التي يمكن استبدالها بغيرها في نفس السياق، لنتأمل ما يلي:

1– ⵉⵙⵍⵎⴰⴷ ⵅⴼ ⵡⴰⵔⵔⴰⵜⵏ ⵏⵏⵙ (اطلع الأستاذ على كتبه)

2-ⵉⵙⵔⵙ ⵉⵙⵍⵎⴰⴷ ⴰⵖⴰⵏⵉⵎ ⵅⴼ ⵡⴰⵔⵔⴰⵜⵏ ⵏⵏⵙ (وضع الأستاذ القلم على كتبه)

من خلال المقارنة بين الجملة (1) والجملة (2) نرى أن أي استبدال للحرف (ⵅⴼ) غير ممكن دون أن يكون تغيير في طبيعة العناصر المشكلة لهذه الجملة، وعلى العكس من ذلك يمكن أن نستبدل الحرف (ⵅⴼ)بمجموعة أخرى من الحروف والظروف، إذ يمكن أن نستبدل (ⵅⴼ) ب(ⴰⴼⵍⵍⴰ) أو (ⴷⴷⴰⵡ) دون تغيير العناصر المشكلة للجملة (2).

لقد أثبتت لنا عملية الاستبدال التي طبقناها على الجمل الثلاثة السابقة أننا أمام نوعين من الأفعال ينتقيان مكملين مختلفين من ناحية الطبيعة الدلالية.

- العائدية (l'anaphoraisation) وهي تشبه عملية الاستبدال، إلا أنها تختلف عنها في إيجاد عنصر لكل نمطية استبدال. وتنحصر دلالتها في طبيعتها الإشارية، والوحدات اللغوية التي تصلح لهذه الوظيفة هي الضمائر والإشاريات، أو ما يسمى بالمعوضات. وتختلف المعوضات الضميرية والإشارية حسب اختلاف طبيعة المكمل كما تبين الجمل (1.1)، (2.1) و(3.1) في الأمثلة التالية:

---

[1] - لمزيد من المعلومات عن محلاتية تنيير يراجع: تنيير (1969)

ولقد استثمرت مجموعة من المجالات الوثيقة الصلة باللغة نتائج التحليل المحلاتي من أهمها مجال المعالجة الآلية للغات و مجال تدريس اللغات الأجنبية.

وتسعى المداخلة المعنونة بـ: نحو حوسبة محلاتية للأفعال في الكتاب المدرسي "ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ" إلى اختبار هذا الإطار انطلاقا من النموذج المحلاتي لإنغل و إبراز الأدوار الوظيفية لتلك المعالجة في مجال تعليم وتعلم اللغة الأمازيغية، انطلاقا من نصوص الكتاب المدرسي.

وستتم مقاربة الموضوع انطلاقا مما يلي:

1. محلاتية الفعل عند إنغل، الحدود والمبادئ، وسنتحدث فيه عن أهم المبادئ التي تقوم عليها محلاتية إنغل مقارنة بالإطارات المحلاتية السابقة لها؛

2. الأدوار الوظيفية و الأبعاد التعليمية للحوسبة المحلاتية للأفعال في الكتاب المدرسي للأمازيغية، وسنشير فيه إلى الانعكاسات التعليمية والأدوار الوظيفية للمعالجة الآلية المحلاتية في تجويد التعلمات في مجال اللغة الأمازيغية؛

3. التطبيقات الحاسوبية للمحلاتية انطلاقا من الكتاب المدرسي للأمازيغية "ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ"

وسنحاول أن نعالج حاسوبيا مقطعا من نص مدرسي مأخوذ من الكتاب المدرسي"ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ" للسنة الثانية باعتماد الإطار النظري لإنغل (1977).

2. محلاتية الفعل عند إنغل، الحدود والمبادئ

تطلق المحلاتية (la valence) في الأدبيات اللسانية المعاصرة، خاصة مع نحو التبعية la grammaire de dépendence، وبعض من الأنحاء الوظيفية Les grammaires fonctionnelles على ذلك النوع من علاقات التبعية التي تضمن الإسناد بين عنصر من عناصر الجملة ، كالفعل مثلا، وبين العناصر التي تدور في مجاله، والتي تختلف عددا ونوعا باختلاف العنصر الذي تأتلف معه (أليرتون،1996: 359)، كما تطلق المحلاتية، أيضا، على تلك القدرة التي يمتلكها الفعل أو الاسم في التأليف مع العناصر الأخرى للجملة.

إن الغرض الأسمى للمحلاتيين هو الوصول إلى تحقيق بحث شامل حول عدد العناصر التي يقتضيها كل فعل، أو اسم في أية لغة من اللغات قيد الدراسة، وكذا تحديد الإمكانات المختلفة لكل فعل على حدة.

نحو حوسبة محلاتية للأفعال في الكتاب المدرسي "ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ" و أبعادها التعليمية

كمال أقا

المعهد الملكي للثقافة الأمازيغية، الرباط، المغرب

aqua@ircam.ma

ملخص

تهدف المداخلة إلى تقديم نموذج أولي لمشروع حوسبة محلاتية لأفعال اللغة الأمازيغية انطلاقا من الكتاب المدرسي (ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ) بالاعتماد على الإطار المحلاتي الذي صاغه إنغل (Engel). إنه تقديم نسعى من ورائه إبراز الأدوار الوظيفية للحوسبة المحلاتية في مجال المعالجة الآلية للمتن الأمازيغي بوجه عام، والمتن المبثوث في الكتاب المدرسي بوجه خاص.وعليه تطرقت المداخلة في البداية للإطار العام لمحلاتية إنغل والمبادئ اللسانية التي يقوم عليها، وأشارت بعد ذلك لأهم انعكاساتها في مجال تعليم وتعلم اللغات، وانتهت المداخلة بتقديم نموذج حوسبة محلاتية لمقطع من نص مدرسي مقتطف من كتاب (ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ) للسنة الثانية باستعمال برنامج معلوماتي.

1. مقدمة

لقد أدى التطور الحاصل في الأبحاث التركيبية ابتداء من منتصف القرن الماضي إلى ظهور مجموعة من الاتجاهات التحليلية التي تعالج اللغات الطبيعية بالاعتماد على مجموعة من المبادئ أهمها: الدقة والبساطة في التحليل، وإمكانية التطبيق، وتوظيف النتائج في مجالات تكون اللغة عنصرها الأساس.ولعل محلاتية الفعل (la valence du verbe) من بين الاتجاهات اللسانية التي انطلقت من هذه المبادئ لمعالجة بعض من قضايا اللغات الطبيعية .

ومحلاتية الفعل هي إطار لساني تركيبي ودلالي يبحث في جهة الأفعال من حيث عدد محلاتها ونوعها وسماتها الدلالية حسب السياقات التي ترد فيها الممكنة والمقبولة تركيبيا دلاليا.

Recognition of Tifinaghe Handwritten Characters using Moments for feature extraction

Mohamed Abaynarh[1], Hakim Elfadili[2], Lahbib Zenkouar[3]

[1]Ecole Mohamadia d'Ingenieurs RABAT

mohamed.abaynarh@gmail.com

[2]Ecole Nationale des Sciences Appliquées, FES

el_fadili_hakim@yahoo.fr

[3]Ecole Mohamadia d'Ingenieurs RABAT

Lahbibzenkouar@emi.ac.ma

Abstract

There has been a significant amount of research in various aspects of writing based user interfaces including interactive design tools, studies of gestures, software toolkits, ink beautification, and sketch recognition. In this paper, we shall focus on the recognition of handwritten characters that are used in common applications.

Different people may use different stroke-order, number, and direction to draw the same shape of any character. In fact, handwritten characters are imprecise in nature such that corners are not always sharp, lines are not perfectly straight, and curves are not necessarily smooth. A robust recognition system has to account for all of these factors. In this paper, we shall consider a statistical approach to Handwritten Character Recognition using Legendre moments as features. These features will be the input of character classification algorithms based on nearest neighbor criteria (NN) (Khotanzad, A. and Hong) and hidden neural network.

Experimentation has been performed on a local database of characters. Experimental results show the robustness of the approach.

Keywords: Character recognition, Tifinaghe characters, Invariant moments, Neural network

Résumé

Nous présentons dans ce manuscrit une méthode de reconnaissance de caractères amazighs manuscrits isolés, basée sur les moments invariants.

La base de données constituée est composée d'images de l'alphabet Tifinaghe écrits par 28 scripteurs pour un total de 924 images de taille normalisée

100x100, Elle est divisée en une base d'apprentissage de 726 caractères et une base de test de 198 caractères.
Des algorithmes d'extraction d'attributs appropriés à base de la méthode des moments serviront à alimenter les classifieurs utilisés : la distance minimale, les plus proches voisins, et les réseaux des neurones.
Les résultats expérimentaux dressent une étude comparative entre les différents algorithmes en termes de taux de reconnaissance

Mots clés : reconnaissance de caractères, caractères Tifinaghe, Moments invariants, Réseaux de neurones

1-Introduction

Character Recognition is a part of Pattern Recognition. It's the research area that studies the operation and design of systems that recognize patterns in data. It encloses sub disciplines like discriminate analysis, feature extraction, error estimation, cluster analysis (also called statistical pattern recognition), grammatical inference and parsing. Important application areas are image analysis, character recognition, speech analysis, man and machine diagnostics, person identification and industrial inspection.

Handwriting is a simple and natural mode of expression. It is especially desirable for conceptual design, both on an Individual basis and in a collaborative environment. It is used in a significant amount of research to date in various aspects of sketch-based user interfaces: Interactive design tools, studies of gestures, software toolkits, ink beautification, and sketch recognition. The work in off-line character recognition can be roughly categorized into statistical, structural, and rule-based approaches.

In this paper, we consider a statistical approach to Tifinaghe handwritten characters recognition using Legendre moments as features. In fact, Legendre moments are a class of orthogonal moments and have been shown effective in terms of image representation.

Legendre moments can be easily constructed to an arbitrary order. Although higher order moments carry more fine details of an image, they are also more susceptible to noise.

Therefore we have experimented with different orders of Legendre moments to determine the optimal order for our problem, for this, we introduce the Maximum Entropy Principle (MEP) as a selection criterion (C.-H. Tech and R.T. Chin, 1988) that extracts optimal character features.

The main goal is to reduce the input dimensionality of the classification problem by eliminating features with low information content or high redundancy with respect to other features.

The second step (recognition) is achieved by using minimum mean distance, multilayer hidden neural network and nearest neighbor as classifier, where finite vectors obtained in the preprocessing phase are used as inputs. Experimental results are obtained using a collected database of handwritten Amazigh characters.

2-Database construction

Since there is no publicly available handwritten database of Tifinaghe characters, we have created a test corpus by gathering data from different and independent people, referring to The alphabet Tifinaghe adopted by IRCAM,(L. Zenkouar,2004,2008) which is composed of thirty-three characters representing consonants and vowels as shown in Figure 1.

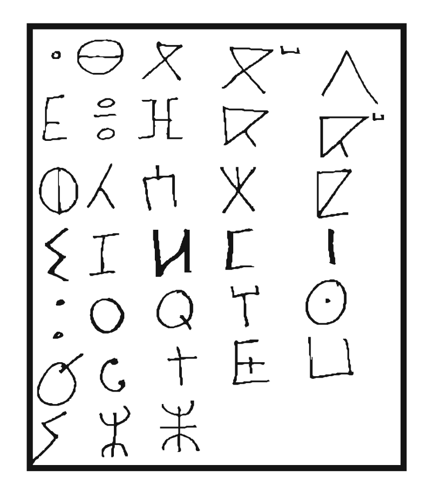


Figure 8 : the characters set representing the Amazigh alphabet adopted by IRCAM

So far, we have gathered data from 28 users. Each user was asked to write one example for each of The 33 characters of Tifinaghe alphabet shown in Figure 1, the resulting dataset contains a total of 28 users overall and 924 characters.

Our database is composed of isolated character images of Tifinaghe alphabet, gathered from 28 users, to obtain 924 character images

The database has the following properties:

- ❖ gray level Images coded with 8 bytes

- All images have 100 X 100 size
- Training and test databases are written by different users, it's divided on training database containing 726 characters, and test database containing 198 characters

Indeed, our database contain 924 gray level character images,

3- Features extraction

In order to design a good character recognition system, the choice of feature extractor is very

crucial. In fact, the feature vectors should contain the most pertinent information about the character to be recognized while having a low dimensionality.

In the statistics-based feature extraction approaches, global information is used to create a set of feature vector elements to perform recognition. The low-dimensional feature vector reduces the computational burden of the recognition system; however, if the choice of the feature elements is not properly made, this in turn may affect the classification performance.

Also, as the number of feature elements in the feature extraction step decreases, the neural network classifier becomes small with a simple structure.

Statistics-based approaches for feature extraction are very important in pattern recognition for their computational efficiency and their use of global information in an image for extracting features (J. Haddadnia and al, 2001). Especially, the advantages of considering orthogonal moments are that they are shift, and scale invariants and are very robust in the presence of noise. The invariant properties of moments are utilized as pattern sensitive features in classification and recognition applications (C. H. Teh and R. T. Chin, 1988; S. O. Belkasim and al 1991).

Statistical moments represent average values of processes (powered to order n) when a

random variable is involved. Here, the original and pre-processed images were considered as

two dimensional arrays of a random variable of dimension $N \times N$. The random variables took values from level 0 to 255, as the images were considered in gray levels quantized in 8 bits

(Gray levels were obtained from BMP format). Moments were calculated for the random variable X, which was identified with the image block. In addition, X is a

matrix of two coordinates *(x,y)* obtained from the image matrix *f(x,y)*. The definition of *(p+q)* order invariant moment around the origin is given by:

The Legendre moments of order $(p + q)$ are defined for a given real image intensity function $f(x, y)$ as

$$\lambda_{p,q} = \frac{(2p+1)(2q+1)}{4} \int_R \int_R P_p(x) P_q(y) f(x,y) dxdy,$$

Where $f(x,y)$ is assumed to have bounded support

The Legendre polynomials $P_p(x)$ are a complete orthogonal basis set on the interval $[-1,1]$ for an order p they are defined as

$$p_p(x) = \frac{1}{2^p p!} \frac{d^p}{dx^p} (x^2 - 1)^p$$

The orthogonality property is guaranteed by the equality

$$\int_{-1}^{1} p_p(x) p_q(x) dx = \frac{2}{(2p+1)} \delta_{p,q}$$

Where $\delta_{p,q}$ is the Kronecker function, that is,

$$\delta_{p,q} = \begin{cases} 1 & \text{if} \quad p = q \\ 0 & \text{otherwise} \end{cases}$$

3-1-Image reconstruction by Legendre moments

By taking the orthogonality principle into consideration, the image function $f(x,y)$ can be written as an infinite series expansion in terms of Legendre polynomials over the square $[-1,1] \times [-1,1]$:

$$f(x,y) = \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \lambda_{p,q} P_p(x) P_q(y).$$

Where the Legendre moments are computed over the same square

If only Legendre moments of order smaller than or equal to θ are given, then the function $f(x,y)$ can be approximated by a continuous function which is a truncated series:

$$f_\theta(x,y) = \sum_{p=0}^{\theta} \sum_{q=0}^{p} \lambda_{p-q} P_{p-q}(x) P_q(y).$$

Furthermore, $\lambda'_{p,q}s$ must be replaced by their numerical approximation which will be pointed out on the following section. The number of moments used in the reconstruction of image for a given θ is defined by

$$N_{total} = \frac{(\theta+1)(\theta+2)}{2}$$

3-2-Approximation of the Legendre moments

In practice the Legendre moments have to be computed from sampled data, that is, the rectangular sampling of the original image function $f(x,y)$, producing the set of samples $f(x_i, y_j)$ with an (M,N) array of pixels, thus we define the discrete version of $\lambda_{p,q}$ in terms of summation by the traditional commonly used formula (C.-H. Tech and R.T. Chin, 1988):

$$\tilde{\lambda}_{p,q} = \frac{(2p+1)(2q+1)}{4} \sum_{i=1}^{M} \sum_{j=1}^{N} P_p(x_i) P_q(y_j) f(x_i, y_j) \Delta x \Delta y$$

Where $\Delta x = (x_i - x_{i-1})$ and $\Delta y = (y_j - y_{j-1})$ are sampling intervals in the x and y directions.

It is clear, however, that $\tilde{\lambda}_{p,q}$ is not a very accurate approximation of $\lambda_{p,q}$, in particular, when the moment order $(p+q)$ increases

The piecewise constant approximation of $f(x,y)$ proposed recently by Liao and Pawlak

(S. X. Liao, 1996; S. X. Liao, 1993), yields the following approximation of $\lambda_{p,q}$:

$$\tilde{\lambda}_{p,q} = \sum_{i=1}^{M} \sum_{j=1}^{N} H_{p,q}(x_i, y_j) f(x_i, y_j).$$

With the supposition that $f(x,y)$ is piecewise constant over the interval

$$[x_i - \Delta x, x_i + \Delta x] \times [y_j - \Delta y, y_j + \Delta y]$$

And where

$$H_{p,q}(x_i, y_j) = \frac{(2p+1)(2q+1)}{4} \int_{x_i - \frac{\Delta x}{2}}^{x_i + \frac{\Delta x}{2}} \int_{y_j - \frac{\Delta y}{2}}^{y_j + \frac{\Delta y}{2}} P_p(x) P_q(y) \, dx \, dy$$

represents the integration of the polynomial $P_p(x)P_q(y)$ around the (x_i, y_j) pixel.

This approximation allows a good quality of reconstructed images by reducing the reconstruction error.

4-Learning and classification

Three different classification techniques have been evaluated: the minimum mean distance (MMD), the nearest neighbor (NN) (Khotanzad, A. and Hong), and hidden neural network. The classifiers learn from the training set in which every example is represented with a multi-dimensional feature vector composed of extracted Legendre moments.

4-1-Minimum Mean Distance (MMD)

In the minimum distance classifier, each character class, C_k, is represented with the sample means, μ_k, learned from the training examples. When a new example is given, it is compared to each character class by calculating the Euclidean distance. The example is assigned to class *k* for which the distance is minimum.

The training example of class *k*, c_k, with the smallest distance to the test example, *a*, is the nearest neighbour of *a*, the equations are shown below:

$$d(a,c^k) = \sum_{i=1}^{m} (a_i - c_i^k)^2$$

4-2-Nearest Neighbour (NN)

During training, the nearest neighbour classifier uses the feature vectors of the samples in the training set using the corresponding moments. In the classification stage, the classifier extracts features from the test example and computes the Euclidean distance, *d*, between the example and every training example.

The test example is classified to the class receiving the maximum number of votes. The training data is scaled to be in the range of [0, 1] in order to avoid numerical problems. The test data is also scaled according to the parameters obtained during the training stage

4-3-Neural Network

Neural network is widely used as a classifier in many recognition systems. Neural networks have been employed and compared to conventional classifiers for a number of classification problems. The results have shown that the accuracy of the neural network approaches is equivalent to, or slightly better than, other methods. Also, due to the simplicity, generality, and good learning ability of the neural networks, these types of classifiers are found to be more efficient (W. Zhou, 1999) Therefore, neural networks are an excellent candidate for pattern classification (J. Haddadnia, and al, 2002), where attempts have been carried out to make the learning process in this type of classification faster than normally required for the multilayer neural networks (W. Zhou, 1999)

In this paper, a neural network is used as a classifier in character recognition where the inputs to the neural network are feature vectors derived from the proposed feature extraction technique described in the previous section (H. El fadili, 2006)

The output of each node is a pondered sum of its inputs:

$$o_i = \varphi(a_i) = \varphi\left(\sum_{k=1}^{N} (w_{ik} \xi_k)\right)$$

with ξ_k the k^{th} composant of sample vector.

W_{ik} is the weight of the connection which rely unit k and unit i.

a_i is the activation of the unit i,

φ is the activation function of the units which is a threshold function with the following expression :

$$\varphi(x)=\begin{cases}+1 & si \quad x\geq\theta \\ -1 & si \quad x\leq\theta\end{cases}$$

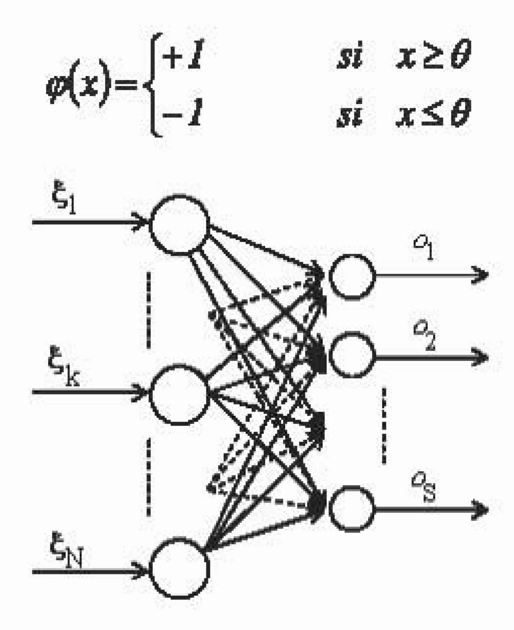


Figure 2: Simple Perceptron

5-Experimentation

We have designed two sets of experiments based on these two usage scenarios to evaluate the recognition system.

In this experiment, we are interested in determining how well the pre-trained recognizer works for a new user under different classification methods. Each time, a different individual's data set is held out for a test set, and a classifier is trained with all other users' data and then test on the holdout set. For each round, there are 726 characters for training, and 198 characters for testing.

MMD (%)	Nearest Neighbour (%)	Neural Network (%)
61	24	9

Table 1: comparison of error rate for the test set of the three used classification methods

Table 1 shows the classifier error rate of three approaches, the classifier error rate (%) is considered as the number of misclassifications in the training (test) phase over the total number of training (test) images.

From Table 1 we can see that neural network method wit, hidden layers (only two hidden layer) and hidden nodes can easily provide excellent results in terms of test error

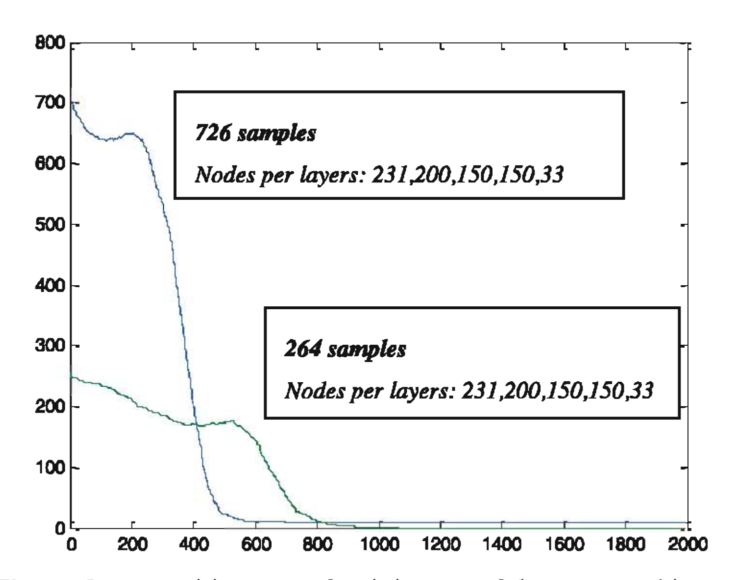


Figure 3: recognition rate of training set of the same architecture

and two different samples

The recognition converges faster when a the number of iterations is great, due to the very small number of training examples

We believe it is because there is a great level of consistency in how a user draws shape (character). Of course, the more examples, the better is to train the recognizer

6-Conclusion

In this paper an efficient feature extraction technique is developed, based on the orthogonal moments using Invariant Legendre moments. We have focused on the discrimination power of Legendre moments and have shown that the proposed

Legendre moment extraction method with hidden neural network classifier is tolerant to shape distortion, while showing improved performances in terms of recognition rate and generalization ability.

Références

C. H. Teh and R. T. Chin, "On image analysis by the methods of moments," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 10, no. 4, pp. 496–513, 1988.

C.-H. Tech and R.T. Chin, "On image analysis by the methods of moments," *IEEE Trans, on Pattern Analysis and Machine Intelligence*, vol. 10, no 4, pp.496-513, 1988.

H. El fadili, "Conception d'un système de reconnaissance de forme par combinaison de la méthode des moments avec un classifieur neuronal optimisé par algorithme d'évolution, " Ph.D. thesis, Laboratoire d'Electronique, Signaux-Systèmes et d'Informatique, University of Sidi Mohamed ben Abdellah, Fes, Morocco, 2006.

J. Haddadnia, K. Faez, and P. Moallem, "Neural network based face recognition with moments invariant," in Proc. *IEEE International Conference on Image Processing (ICIP '01),* vol. 1, pp. 1018–1021, Thessaloniki, Greece, October 2001.

J. Haddadnia, M. Ahmadi, and K. Faez, "An efficient method for recognition of human faces using higher orders pseudo Zernike moment invariant," in *Proc. 5th International Conference On Automatic Face and Gesture Recognition (FG '02),* pp. 330–335,Washington, DC, USA, May 2002.

J. Haddadnia, K. Faez, and P. Moallem, "Neural network based face recognition with moments invariant," in Proc. *IEEE International Conference on Image Processing (ICIP '01),* vol. 1, pp. 1018–1021, Thessaloniki, Greece, October 2001.

J. Haddadnia, M. Ahmadi, K. Faez"An Efficient Feature Extraction Method with Pseudo-Zernike Moment in RBF Neural Network-Based Human Face Recognition System" *EURASIP Journal on Applied Signal Processing*, 890–901 2003 Hindawi Publishing Corporation

J. Haddadnia, M. Ahmadi, and K. Faez, "A hybrid learning RBF neural network for human face recognition with pseudo Zernike moment invariant," in *IEEE International Joint Conference On Neural Network (IJCNN '02),* pp. 11–16, Honolulu, Hawaii, USA, May 2002.

Khotanzad, A. and Hong, Y.H. Invariant Image Recognition by Zernike Moments. *IEEE Trans. on PAMI,* 12(5). 289-497.

L. Yingwei, N. Sundararajan, and P. Saratchandran, "Performance evaluation of a sequential minimal radial basis function (RBF) neural network learning algorithm," *IEEE Transactions on Neural Networks,* vol. 9, no. 2, pp. 308–318, 1998.

L. Zenkouar, « L'écriture Amazighe Tifinaghe et Unicode », *in Etudes et documents berbères.* Paris (France). n° 22, pp. 175—192, 2004

L. Zenkouar, « Normes des technologies de l'information pour l'ancrage de l'écriture amazighe », *revue Etudes et Documents Berbères,* Paris(France), n°27, pp. 159-172, 2008

S. O. Belkasim, M. Shridhar, and M. Ahmadi, "Pattern recognition with moment invariants: a comparative study and new results," *Pattern Recognition,* vol. 24, no. 12, pp. 1117–1138, 1991.

Stone, M. Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Royal Statist. Soc.* B, 36. 111-147.

S. X. Liao and M. Pawlak, "On image analysis by moments," *IEEE Trans, on Pattern Analysis and Machine Intelligence,* vol. 18,no. 3, pp.254-266,1996.

S. X. Liao, Image analysis by moments, Ph.D. thesis, Department of Electrical and computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada, 1993.

W. Zhou, "Verification of the nonparametric characteristics of backpropagation neural networks for image classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 37,no. 2, pp. 771–779, 1999

Reconnaissance Automatique de l'Ecriture Amazighe à base de Ligne Centrale de l'Écriture

Youssef Es Saady, Ali Rachidi, Mostafa El Yassa, Driss Mammass

Laboratoire IRF-SIC, Université Ibn Zohr B.P. 8106, Hay Dakhla Agadir, Maroc
essaady2110@yahoo.fr , rachidi.ali@menara.ma, melyass@gmail.com, mammass@univ-ibnzohr.ac.ma

Résumé

Nous présentons dans ce papier un système automatique de reconnaissance de l'écriture amazighe à base de ligne centrale de l'écriture. Après des prétraitements sur l'image, le texte est segmenté en lignes et puis en caractères. Les positions des lignes de base du caractère sont utilisées pour obtenir un ensemble de caractéristiques indépendantes et dépendantes à ces lignes. Ces caractéristiques sont liées aux densités de pixels et sont extraites sur les images binaires des caractères en se basant sur l'utilisation de la technique des fenêtres glissantes. Ces primitives alimenteront un réseau de neurones multicouches dans les phases d'apprentissage et de reconnaissance. Le système a montré de bonnes performances sur une base de 19437 modèles amazighes.

Mots clefs : Reconnaissance d'écriture, Caractères amazighs, Ligne de base, Segmentation, Perceptron multicouches.

1. Introduction

La reconnaissance automatique de l'écriture manuscrite ou imprimée reste encore un sujet de recherche et d'expérimentation. Le problème n'est pas encore entièrement résolu bien que les résultats atteignent des taux assez élevés dans certaines applications et pour certaines langues. Plusieurs recherches scientifiques ont été effectuées sur l'écriture latine, arabe, et autres, ce qui a permis le

développement de plusieurs approches de reconnaissance automatique de ces écritures. Par contre, l'écriture amazighe, appelée Tifinaghe, est très peu traitée. Quelques tentatives ont été menées pour améliorer la situation actuelle. Elles sont regroupées généralement en grandes classes telles que les approches statistiques (Oulamara, 1988), (Djematen et al., 1997), Les réseaux de neurones (Ait Ouguengay, 2008), (Elyachi et al., 2009), (Bouikhalene et al., 2009), l'approche syntaxique (Es Saady et al., 2008), (Es Saady et al., 2010) et les Modèles de Markov cachés (Amrouch et al., 2009), (Amrouch et al., 2010). Dans ce cadre, nous avons réalisé un système automatique de reconnaissance de caractères amazighes imprimés isolés, basé sur une approche syntaxique utilisant les automates finis. Sur une base de données de caractères amazighes imprimés segmentés isolés, des résultats encourageants ont été obtenus sur la majorité des caractères. La limite de cette approche est qu'elle n'est pas applicable pour les caractères non segmentés. Et pour remédier à ces limites, on propose une nouvelle approche qui tient compte de tous les caractères amazighes.

En effet, dans la phase d'extraction des primitives, notre approche est basée sur la position de la ligne de base d'écriture. Ces primitives alimenteront un réseau de neurones multicouches dans les phases d'apprentissage et de reconnaissance.

Pour les écritures des autres langues, différentes approches basées sur les positions des lignes d'écriture ont été proposées dans la littérature. Pour l'écriture latine et arabe, plusieurs positions de lignes on été utilisées pour extraire des caractéristiques qui dépendent de ces lignes (Elhajj et al., 2005), (AL-Shatnawi and Khairuddin, 2008), (Aida-zade and Hasanov, 2009), (Razzak et al., 2010). La figure 1 illustre des exemples de lignes utilisées pour les caractères latins et les caractères arabes (Elhajj et al., 2005). Dans le cas de l'écriture amazighe, on propose d'utiliser une ligne centrale, ligne supérieure et inférieure de l'écriture pour dériver un ensemble de caractéristiques indépendantes et dépendantes à ces lignes. Ces caractéristiques sont de types statistiques extraits au niveau pixels en se basant sur l'utilisation de la technique des fenêtres glissantes.

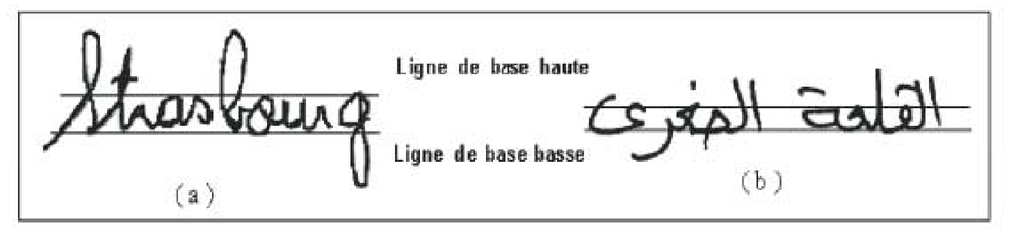


Figure 1: Exemple de lignes de base d'écriture. (a) cas de l'écriture latine, (b) cas de l'écriture arabe.

L'architecture générale de notre système de reconnaissance de caractères amazighs se présentera dans la figure 2 ci-dessous.

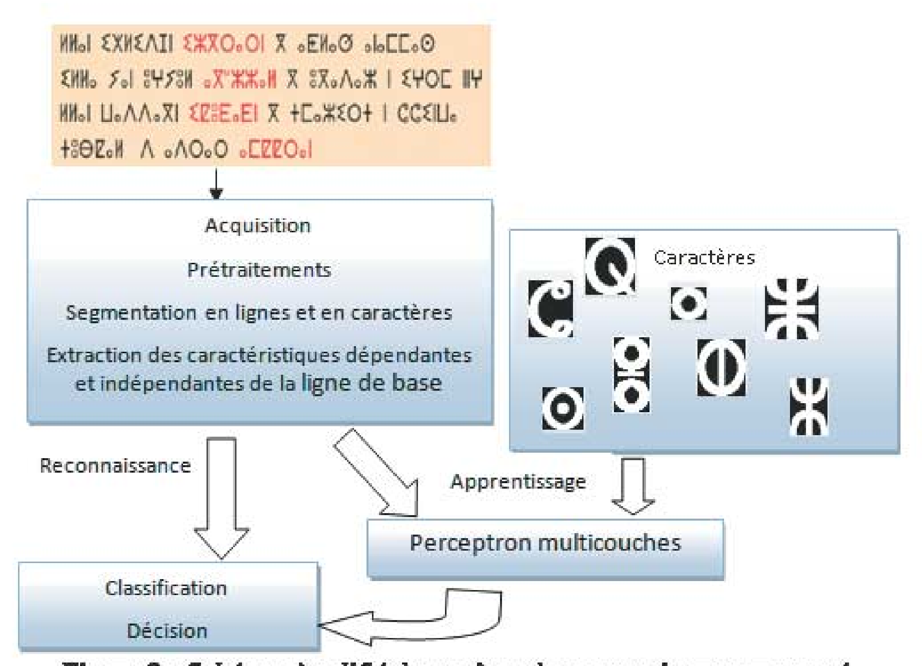


Figure 2 : Schéma simplifié du système de reconnaissance proposé.

Dans la section 2, nous présentons les principales caractéristiques de l'écriture amazighe. La troisième section est consacrée à la présentation des étapes de prétraitements effectuées dans le système. Ensuite, nous présenterons dans la section 4 les caractéristiques extraites en utilisant la technique des fenêtres verticales glissantes et les différentes lignes de base. Les résultats expérimentaux

seront présentés et commentés dans la section 5. Finalement, une conclusion ainsi que des perspectives futures sont présentées dans la section 6.

2. Ecriture Amazighe

Le Tifinaghe est le système d'écriture de la langue amazighe. Il tire son origine du vieil alphabet libyque et saharien, déjà utilisé depuis le VI^{ème} siècle avant l'ère chrétienne par les populations de l'Afrique du Nord, du Sahel et des Iles Canaries. Cet alphabet a subi des modifications et des variations depuis son origine jusqu'à nos jours.

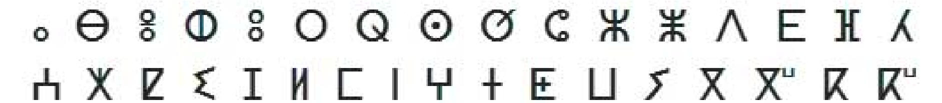


Figure 3 : Alphabet Tifinaghe IRCAM.

La figure 3 ci-dessus présente les différents modèles de l'alphabet amazighe (Tifinaghe-IRCAM). Il comporte trente trois lettres. A la différence des caractères latins et arabes, l'écriture amazighe n'est jamais cursive, ce qui facilite toute opération de segmentation. La majorité des modèles graphiques des caractères est composée de points, de petits cercles, et/ou de segments. De plus, l'écriture amazighe est écrite de gauche à droite, elle utilise des signes de ponctuation classique acceptés en alphabet latin. La figure 4 ci-dessous présente un exemple du texte amazigh dans un manuel scolaire.

ⴰⴷ ⵙⵎⴷⵢ ⵜⵉⵡⵉⵏⴰⵙ

ⵉⵎⴷⵢⴰⵜⵏ :

1- ⵜⴰⵖⴰⵄⵜ ⵉⵏⵓ, ⵖⵓⵔⵙ ⵜⴰⵙⴽⵉⵡⵉⵏ ⵜⵓⵙⴷⵉⴷⵉⵏ.
2- ⵍⵍⴰⵏ ⵉⴳⵍⵉⴷⵏⵏ ⵉⵣⴳⵔⴰⵔⵏ ⴳ ⴰⵃⵍⴰⵚ ⴰⵍⴰⵎⵎⴰⵙ.
3- ⵉⵍⵍⴰ ⵢⴰⵏ ⵓⵖⵢⵓⵍ ⴰⴳⵯⵣⵣⴰⵍ ⴳ ⵓⴳⴰⴷⴰⵣ ⵏ ⵉⵖⵔⵎ ⵏⵏⵖ.
4- ⵍⵍⴰⵏ ⵡⴰⴷⴷⴰⴳⵏ ⵉⵇⵓⵄⴰⵄⵏ ⴳ ⵜⵎⴰⵣⵉⵔⵜ ⵏ ⵛⵛⵉⵍⵡⴰ.
5- ⵜⵓⴱⵇⴰⵍ, ⴷ ⴰⴷⵔⴰⵔ ⴰⵎⵇⵇⵔⴰⵏ.

Figure 4: Exemple du texte Amazighe dans un manuel scolaire.

3. Prétraitements

Cette étape prépare l'image d'entrée afin de faciliter l'étape d'extraction des caractéristiques. Il s'agit essentiellement de réduire le bruit superposé aux textes et essayer de ne garder que l'information significative de la forme représentée. Une fois l'image est numérisée, une série de prétraitements est appliquée. Nous avons utilisé le seuillage, la réduction du bruit, la segmentation en lignes puis en caractères, et enfin la normalisation en taille.

La séparation Avant/Arrière plan est réalisée avec une binarisation. Il s'agit de passer d'une image en niveau de gris ou en couleur à une image bitonale (noir et blanc) en se basant sur un seuil global. Nous avons utilisé la méthode d'Otsu pour la binarisation (Otsu, 1979). La figure 5 ci-dessous présente le résultat de binarisation obtenu avec la méthode d'Otsu. Cette méthode effectue une analyse statistique sur les histogrammes (variance intra-classe et variance inter-classe) pour définir une fonction à maximiser qui permette d'estimer le seuil.

Figure 5: Binarisation avec la méthode d'Otsu.

Pour la réduction de bruit qui consiste à détecter et à éliminer les pixels qui représentent des bruits. Plusieurs méthodes ont été utilisées pour éliminer le bruit. Nous avons utilisé le lissage pour remplacer la valeur d'un pixel par la moyenne des valeurs des pixels entourant (et incluant) le pixel d'origine (Kharma and Ward, 1999).

Une fois l'image du texte est nettoyée, le texte sera segmenté en lignes. Nous avons utilisé les techniques d'analyse d'histogramme de projections horizontales des pixels de manière à distinguer les régions de forte densité (les lignes) des régions de faible densité (les espaces inter-lignes) (cf. figure 6). Ces techniques ont été utilisées souvent pour extraire des lignes dans les textes imprimés, qui ne présentent pas autant de variabilité au niveau de la disposition spatiale des entités connexes comme l'écriture amazighe imprimée.

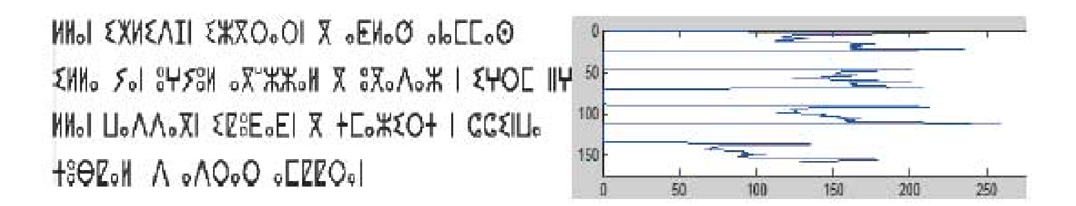


Figure 6: Histogramme de projections horizontales.

L'écriture amazighe n'est pas cursive, cela facilite l'opération de segmentation d'une ligne de texte en caractères. Nous avons utilisé l'histogramme de projections verticales pour segmenter chaque ligne de texte en caractères. La figure 7 ci-dessous présente une ligne de texte, son l'histogramme vertical et le résultat de la segmentation en caractères.

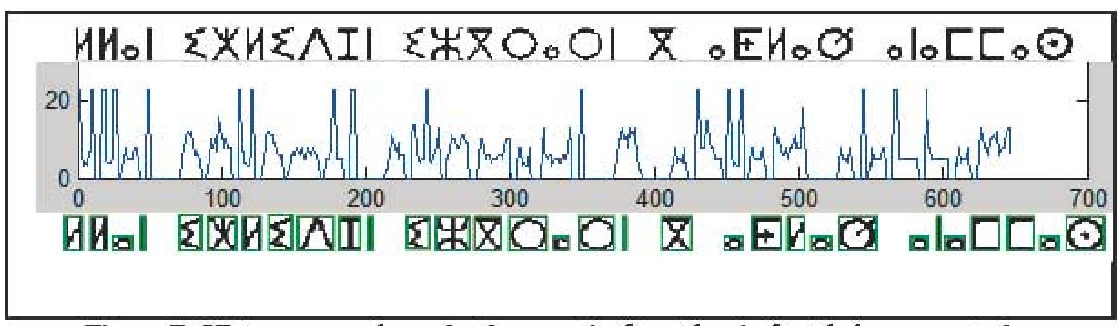


Figure 7: Histogramme de projections verticales et le résultat de la segmentation en caractères.

4. Extraction des caractéristiques

L'étape d'extraction des caractéristiques est précédée d'une étape de prétraitement qui permet d'extraire la ligne de base du caractère. Cette étape permet de séparer l'image du caractère en deux zones : une zone supérieure qui correspond à la zone en dessus de la ligne de base, qui est la line centrale, et une zone inférieure qui correspond à la zone en dessous de la ligne de base. La figure 8 ci-dessous montre les positions des lignes d'écriture sur quelques caractères Amazighes.

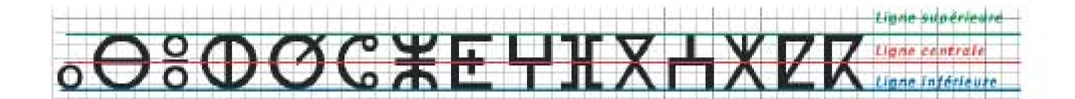


Figure 8: Les positions des lignes d'écriture sur quelques caractères Amazighes.

L'image du caractère est ensuite balayée de gauche à droite et de haut en bas par une fenêtre glissante qui s'adapte en hauteur à celle du caractère (Elhajj et al., 2005). Ces fenêtres sont de largeur fixe (cf. Figure 9).

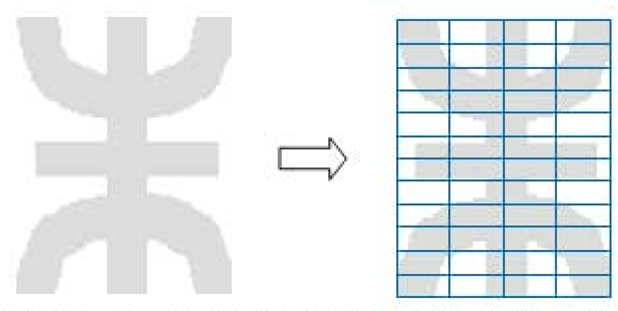


Figure 9: L'image du caractère est divisée en fenêtres verticales.

Dans chaque fenêtre on génère un ensemble de 19 caractéristiques. Celles ci sont représentatives des densités des pixels d'écriture. Pour cela, chaque fenêtre est divisée en un nombre de cellules fixe. Un sous ensemble des caractéristiques sont liées à la position de la ligne de base pour prendre en compte la liaison de la majorité des caractères amazighes par cette ligne.

Nous appelons L la position (ordonnée y) de la ligne de base, qui est la ligne centrale du caractère. Supposons que H est la hauteur en pixels de la fenêtre dans chaque image, h est la hauteur de chaque cellule et W est la largeur de chaque fenêtre. La fenêtre étant divisée verticalement en n_c cellules, donc n_c=H/h. Soit :

- n(i): Le nombre de pixels d'écriture (pixels noirs) dans la cellule i.
- r(j): Le nombre de pixels d'écriture dans la $j^{ème}$ rangée de pixels dans une fenêtre verticale (une fenêtre contient H rangées de pixels).
- b(i): Le niveau d'intensité de la cellule i: b(i)=0 si n(i)=0, b(i)=1 sinon.

Les caractéristiques de densités sont les suivantes:

f_1: densité des pixels noirs dans la fenêtre $$f_1 = \frac{1}{H+w}\sum_{i=1}^{n_c} n(i)$$

f_2: nombre de transitions Noir/Blanc entre cellules $$f_2 = \sum_{i=2}^{n_c} \left| b(i) - b(i-1) \right|$$

f_3: différence de position entre les centres de gravité g des pixels d'écriture dans deux fenêtres consécutives (l'indice *t* est omis) :

$$f_3 = g(t) - g(t-1)$$ Où la position g est calculée comme suit :

$$g = \frac{\sum_{j=1}^{H} j.r(j)}{\sum_{j=1}^{H} r(j)}$$

f_4 à f_{13}: sont les densités de pixels d'écriture dans chaque colonne de la fenêtre.

Les caractéristiques suivantes dépendent de la position de la ligne de base.

f_{14}: position verticale normalisée du centre de gravité des pixels d'écriture, par rapport à la ligne de base.

$$f_{14} = \frac{g-L}{H}$$ Avec L est la position de la ligne de base.

$f_{15} - f_{16}$: deux primitives qui représentent les densités des pixels d'écriture au dessus et au dessous de la ligne de base.

$$f_{15} = \frac{\sum_{j=L+1}^{H} r(j)}{H.w} \quad , \quad f_{16} = \frac{\sum_{j=1}^{L-1} r(j)}{H.w}$$

$f_{17} - f_{18}$: nombre de transitions Noir/Blanc entre les cellules situées au dessus et au dessous de la ligne de base.

$$f_{17} = \sum_{i=k}^{n_c} \left| b(i) - b(i-1) \right| \quad , \quad f_{18} = \sum_{i=2}^{k} \left| b(i) - b(i-1) \right|$$

où k est la cellule contenant la ligne de base.

f_{19}: densité des pixels noirs dans la ligne de base.

L'ensemble des 19 caractéristiques extraites comporte 6 caractéristiques qui dépendent de la position de la ligne de base, et 13 qui n'en dépendent pas.

5. Apprentissage, Expérimentations et Résultats

5.1. Base de données utilisée

Pour évaluer la performance de la méthode proposée, des expériences ont été réalisées sur une base des modèles de la graphie amazighe élaborée par Ait Ouguengay (Ait Ouguengay 2006). C'est une base des modèles de différentes fontes amazighes et de tailles variées. Elle contient au total 12 polices de caractères et les tailles du 10 points au 28 points pour chaque modèle.

Les modèles sont fournis sous forme d'images bitonales de tailles variables. La taille maximale est de 102×129 pixels, tandis que la taille minimale est de 19×2 pixels. Une telle disparité s'explique par le fait que le caractère '∘' (ya) est un petit cercle, et est donc beaucoup plus petit que les autres caractères. Outre le cas particulier du caractère ya, la base est constituée des patterns de différentes fontes amazighes et de tailles variées, qui ne sont pas normalisées.

La manière dont sont stockées les images des modèles, dans cette base, ne permet pas la possibilité de re-normaliser leur taille en une taille moyenne fixe. En effet, Ceci peut être gênant en particulier à cause de la ressemblance des caractères '∘' (ya) et 'O' (yar), qui ne se différencient que par la taille: le caractère ya est un petit cercle, tandis que le caractère yar est un grand cercle. Dans certains cas, on aura une confusion réelle entre des images de ces deux classes. Ce problème aura une influence sur les résultats. Chose qui va être traitée dans les futurs travaux en essayant de développer une base de caractères plus sophistiquée.

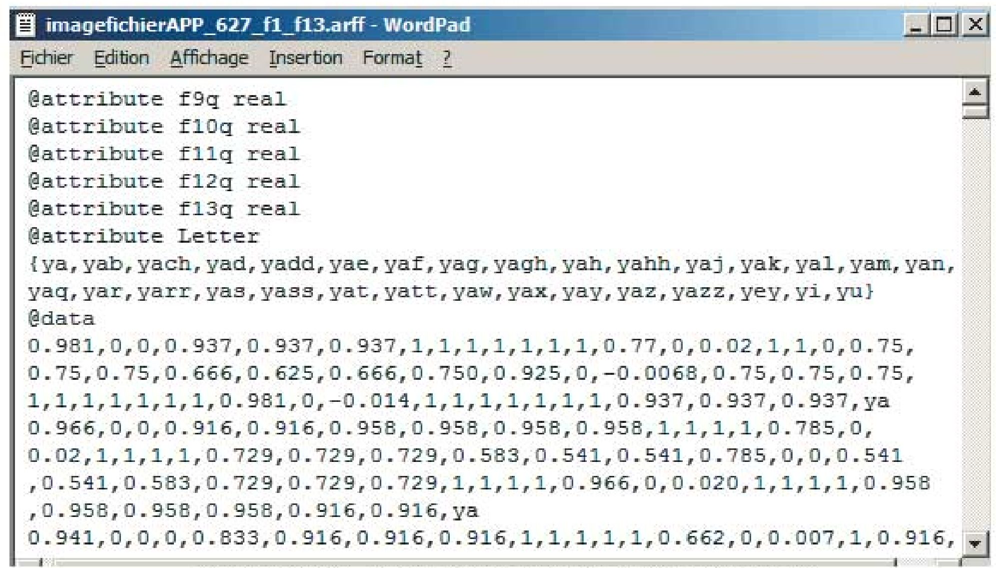
Notre système exige que les images des caractères en entrée soit d'une taille normalisée. Pour cela, nous avons normalisé ces caractères en une taille moyenne 48×40. Ces images de taille normalisée et en format prétraité seront directement soumises en entrée au module d'extraction des caractéristiques.

5.2. Expérimentations et Résultats

De nombreux algorithmes de classification automatique existent et plusieurs implémentations de ces derniers sont disponibles au téléchargement. Plusieurs boites à outils d'apprentissage regroupent de telles implémentations, ce qui en fait des outils idéaux pour lancer des expériences systématiques. Nous avons fait le choix de la plate-forme Weka (Witten et Frank, 2005) pour réaliser l'apprentissage et testons la méthode proposée.

WEKA est un projet open source de l'Université de Waikato (Witten et Frank, 2005). Il a été largement utilisé dans les universités et par plusieurs chercheurs du monde dans le domaine d'exploration de données. Cet outil public propose un ensemble varié d'algorithmes d'apprentissage prêts à l'emploi pour la fouille de donnée. Nous utilisons la méthode de classification: réseaux de neurones, perceptrons multi couches (Multi Layer Perception (MLP). Le perceptron multicouche de WEKA (MLP) a été mis en œuvre par Malcolm Ware en 2000 (Ware, 2000). Son utilisation a été documentée dans un certain nombre de publications de recherche (Klautau 2002). Nous avons utilisé ce classifieur avec ses paramètres par défaut.

De plus, Weka spécifie un format standard aux fichiers d'entraînement et de test (ce sont des fichiers texte avec une extension *.arff). Pour cela nous avons généré deux fichiers, un pour l'apprentissage et l'autre pour le test. La figure 10 présente un extrait d'un exemple de fichier d'entraînement.




```
@attribute f9q real
@attribute f10q real
@attribute f11q real
@attribute f12q real
@attribute f13q real
@attribute Letter
{ya,yab,yach,yad,yadd,yae,yaf,yag,yagh,yah,yahh,yaj,yak,yal,yam,yan,
yaq,yar,yarr,yas,yass,yat,yatt,yaw,yax,yay,yaz,yazz,yey,yi,yu}
@data
0.981,0,0,0.937,0.937,0.937,1,1,1,1,1,1,1,0.77,0,0.02,1,1,0,0.75,
0.75,0.75,0.666,0.625,0.666,0.750,0.925,0,-0.0068,0.75,0.75,0.75,
1,1,1,1,1,1,1,0.981,0,-0.014,1,1,1,1,1,1,1,0.937,0.937,0.937,ya
0.966,0,0,0.916,0.916,0.958,0.958,0.958,0.958,1,1,1,1,0.785,0,
0.02,1,1,1,1,0.729,0.729,0.729,0.583,0.541,0.541,0.785,0,0,0.541
,0.541,0.583,0.729,0.729,0.729,1,1,1,1,0.966,0,0.020,1,1,1,1,0.958
,0.958,0.958,0.958,0.916,0.916,ya
0.941,0,0,0,0.833,0.916,0.916,0.916,1,1,1,1,1,0.662,0,0.007,1,0.916,
```

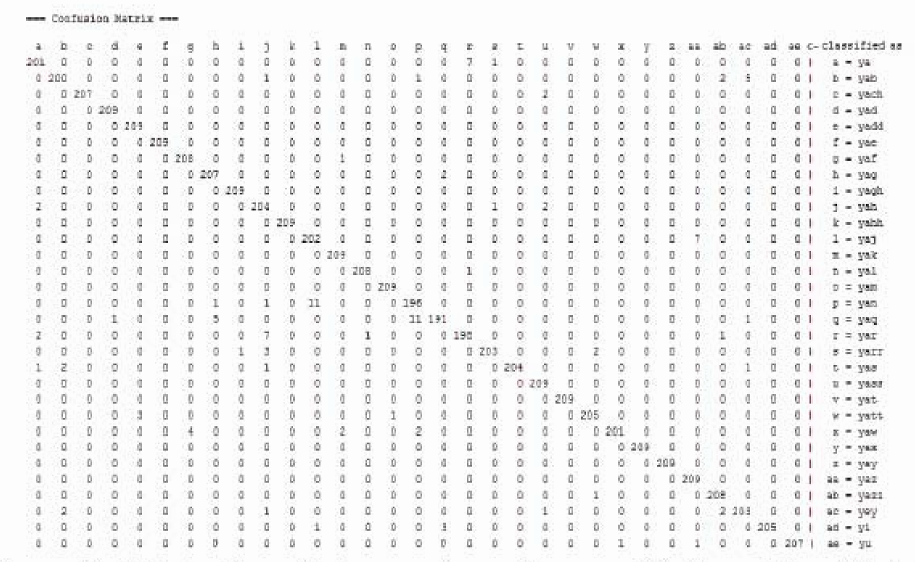
Figure 10: Un extrait du fichier .arff d'entraînement.

Pour évaluer les performances de la méthode proposée, des expériences ont été réalisées sur la base des caractères amazighe décrit précédemment. Les tests ont été effectués en fonction de l'intégration des caractéristiques, dépendantes et indépendantes de la ligne de base. Ainsi, trois expériences ont été réalisées sur un ensemble de 19437 exemples (31 × 627) : un sous ensemble de 12958 images (66,67%) pour l'apprentissage, et un sous ensemble de 6479 images (33,33%) pour le test. Les deux classes sont équiprobables.

Le tableau 1 ci-dessous, présente les résultats du système proposé. Le taux de reconnaissance atteint 98,25 % lorsqu'on intègre les caractéristiques basées sur la position de la ligne de base. Ce qui montre que les caractéristiques basées sur la position de la ligne de base offrent une amélioration significative aux performances de reconnaissance.

Les caractéristiques intégrées	**Apprentissage**		**Test**	
	Taille	*Taux d'appr.*	*Taille*	*Taux de Recon.*
$f_1,\ldots, f13$ (indépendantes de la ligne de base)	12958	**88,78%**	6479	**85,38%**
$f_{14}, \ldots, f_{19}$ (dépendantes de la ligne de base)	12958	**95,89%**	6479	**94,67%**
$f_1, .., f_{19}$ (dépendantes et indépendantes de la ligne de base)	12958	**98,98%**	6479	**98,25%**

Table 1 : Résultats de reconnaissance en fonction des caractéristiques intégrées.



=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	ab	ac	ad	ae	<-- classified as
201	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	1	0	0	0	0	0	0	0	0	0	0	0	0	a = ya
0	200	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	5	0	0	b = yab
0	0	207	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	c = yach
0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	d = yad
0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	e = yadd
0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	f = yae
0	0	0	0	0	0	208	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	g = yaf
0	0	0	0	0	0	0	207	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	h = yag
0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	i = yagh
2	0	0	0	0	0	0	0	0	204	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	j = yah
0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	k = yahh
0	0	0	0	0	0	0	0	0	0	0	202	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	l = yaj
0	0	0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	m = yak
0	0	0	0	0	0	0	0	0	0	0	0	0	208	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	n = yal
0	0	0	0	0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	o = yam
0	0	0	0	0	0	0	1	0	1	0	11	0	0	0	196	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p = yan
0	0	0	1	0	0	0	5	0	0	0	0	0	0	0	11	191	0	0	0	0	0	0	0	0	0	0	0	1	0	0	q = yaq
2	0	0	0	0	0	0	0	0	7	0	0	0	1	0	0	0	198	0	0	0	0	0	0	0	0	0	1	0	0	0	r = yar
0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	203	0	0	0	2	0	0	0	0	0	0	0	0	s = yarr
1	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	204	0	0	0	0	0	0	0	0	1	0	0	t = yas
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	0	u = yass
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	0	0	0	v = yat
0	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	205	0	0	0	0	0	0	0	0	w = yatt
0	0	0	0	0	0	4	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	201	0	0	0	0	0	0	0	x = yaw
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	0	0	y = yax
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	0	z = yay
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	209	0	0	0	0	aa = yaz
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	208	0	0	0	ab = yazz
0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	203	0	0	ac = yey
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	205	0	ad = yi
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	207	ae = yu

Figure 11: Matrice de confusion pour le cas des caractéristiques f1, .., f19, testé sur la base de 6479 caractères.

Les causes d'erreurs sont principalement dues à la ressemblance entre certains caractères sur des fontes différentes. En effet, l'étude de la matrice de confusion du système sur la base des tests, présentée sur la figure 11 ci-dessus, a mis en évidence que la majorité des erreurs étaient faites sur les caractères yaq (X), yan (I), yar (O), yab (Θ), ya (∘), yaw (Ц). A titre d'exemple, 11 images du caractère yan (I) on étés reconnus comme caractère yaj (I). D'ailleurs, le format du caractère yan (I) sur la fonte ' tassafut' ressemble entièrement au caractère yaj (I), comme le montre la figure 12 ci-dessous.

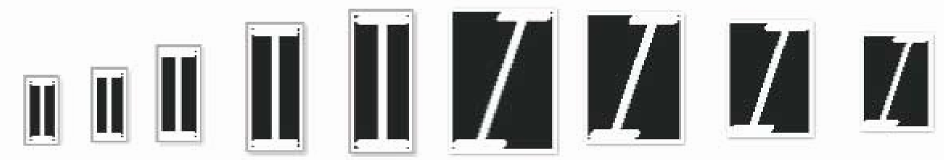


Figure 12: Quelques exemples du caractère yan (I) dans la base, dont la fonte est ' tassafut'

6. Conclusion et Perspectives

Dans cet article, nous avons présenté un système pour la reconnaissance automatique de l'écriture amazighe à base de la position de la ligne de base de chaque caractère. Plusieurs caractéristiques ont été étudiées et comparées. L'importance de l'utilisation de la position de la ligne de base dans l'image du caractère a été prouvée. Les caractéristiques extraites sont basées sur la densité des pixels de dérivée dans une fenêtre glissante. Le système développé a été expérimenté sur une base des modèles de la graphie amazighe. Les résultats montrent une amélioration significative du taux de reconnaissance lorsqu'on intègre les caractéristiques dépendantes de la ligne de base. Parmi les travaux futurs de ce travail, nous allons ajouter d'autres caractéristiques qui améliorent les résultats pour certains caractères dont le taux de reconnaissance est faible par apport aux restes. En plus, nous allons appliquer notre approche sur une base de données manuscrite qui va être développée localement.

Références

Aida-zade R. and Hasanov Z. (2009). Word base line detection in handwritten text recognition systems, International Journal of Electrical and Computer Engineering 4:5 2009, pp. 310-314.

Ait Ouguengay Y., Taalabi M. (2008). Elaboration d'un réseau de neurones artificiel pour la reconnaissance optique de la graphie amazighe, Phase d'apprentissage, SITA'08, INPT, Maroc.

AL-Shatnawi A., Khairuddin O. (2008). Methods of Arabic Language Baseline Detection – The State of Art, IJCSNS International Journal, Vol.8 No.10, pp. 137-143.

Amrouch M., Es saady Y., Rachidi A., Elyassa M., Mammass D. (2009). Printed Amazigh Character Recognition by a Hybrid Approach Based on Hidden Markov Models and the Hough Transform, International Conference on Multimedia Computing and Systems, ICMCS'09, Ouarzazate, Maroc.

Amrouch M., Rachidi A., El Yassa M., Mammass D. (2010). Handwritten Amazigh Character Recognition Based On Hidden Markov Models, ICGST-GVIP Journal, Vol 10, Issue 5: 11-18.

Bouikhalene, M.Fakir B., Safi S., El Kessab B. (2009). Reconnaissance des Caractères Tifinaghe Par L'utilisation Des Réseaux de Neurones Multicouches, SITACAM'09, Agadir, Maroc.

Djematen A., Taconet B., Zahour A. (1997). A Geometrical Method for Printing and Handwritten Berber Character Recognition. ICDAR, pp. 564, ICDAR'97.

Djematen A., Taconet B., Zahour A. (1998). Une méthode statistique pour la reconnaissance de caractères berbères manuscrits; CIFED'98, pp. 170-178.

El-Hajj R., Likforman-Sulem L., Mokbel C. (2005). Arabic handwriting recognition using baseline dependent features and Hidden Markov Modeling, ICDAR 05, Seoul, Corée du Sud.

Elyachi R. and Fakir M. (2009). Recognition of Tifinaghe Characters using Neural Network, International Conference on Multimedia Computing and Systems, ICMCS'09, Ouarzazate, Maroc.

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2010). Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata, ICGST-GVIP Journal, Vol 10, Issue 2: 1-8.

Es Saady Y., Rachidi A., Elyassa M., Mammass D. (2008). Une méthode syntaxique pour la reconnaissance de caractères Amazighes imprimés, CARI'08- Maroc.

Kharma N. and Ward R.K. (1999). Systèmes de reconnaissance de caractères pour les non-experts. IEEE Canadian Review – Summer.

Klautau, A. (2002). Classification of peterson and barney's vowels using weka. Technical report, Universidade Federal do Par.

Otsu N. (1979). A threshold selection method from gray-level histograms, IEEE Trans. Sys,Man., Cyber, vol. 9, pp. 62–66.

Oulamara A., Duvernoy J. (1988). An application of the Hough transform to automatic recognition of Berber characters. Signal Processing, vol. 14: 79-90.

Razzak M., Sher M., Hussain S. (2010). Locally baseline detection for online Arabic script based languages character recognition, International Journal of the Physical Sciences Vol. 5(7), pp. 955-959.

Ware, M. (2000). WEKA Documentation. University of Waikoto.

Witten, I. H., and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers, second edition.

Reconnaissance des Caractères Amazighs par les Modèles de Markov cachés

Brahim Bazdouz, Mohammed Fakir, Belaid Bouikhalene

Equipe de traitement de l'information et de télécommunications, Faculté des Sciences et Techniques, USMS, Maroc
{Bazdous, fakfad, bbouikhalene}@yahoo.fr

1. Introduction

Dans le domaine de la reconnaissance automatique des caractères, plusieurs recherches scientifiques ont été effectuées sur les caractères latins, arabes, et autres. Ceci a permis le développement de plusieurs approches de reconnaissance automatique pour ces caractères. Par contre, les caractères Amazighe, appelés Tifinaghe, sont très peu traités. Et pour extraire les informations Amazighes sur des supports, la reconnaissance automatique est devenue primordiale (Elkessab, 2009).

Ce travail consiste à reconnaître les caractères amazighs manuscrits. Le traitement de chaque caractère commence par les prétraitements afin d'enlever toute sorte de bruits liés à la phase d'acquisition. Ensuite, par l'extraction d'une information sur les différentes directions de son tracé de base. Cette information est exploitée pour générer une séquence d'observations. La séquence obtenue est utilisée pour entraîner un Modèle de Markov Caché pour chaque caractère. L'apprentissage est réalisé avec l'algorithme de Baum-Welch. La classification est enfin effectuée par recherche du modèle discriminant.

L'utilisation des HMMs en reconnaissance automatique de l'écrit a permis d'obtenir des résultats intéressants pour certaines applications grâce à leur capacité d'intégration du contexte et d'absorption du bruit (Pechwitz et al, 2000). Les différents travaux réalisés reposent pour une grande partie sur l'expérience accumulée dans le domaine de la reconnaissance de la parole où les HMMs sont fréquemment utilisés. Comparés à d'autres approches de reconnaissance (structurelle, géométrique, etc.), les HMMs se distinguent par leur capacité de modéliser efficacement différentes sources de connaissance. En effet, d' une part

ils offrent une intégration cohérente de différents niveaux de modélisation (morphologique, lexicale et syntaxique) et d' autre part, il existe des algorithmes puissants permettant de déterminer la valeur optimale des paramètres fournissant la meilleure adéquation entre le modèle et la base de données (connue) qualifiée d'apprentissage.

Dans ce papier, nous présentons une méthode de reconnaissance des caractères Tifinaghe en se basant sur les modèles de Markov Cachés. Les étapes du système développé sont illustrées dans la figure 1.

L'organisation de cet article est comme suit. Dans la deuxième section nous présentons un résumé de la théorie markovienne, dans la troisième section nous rappelons les principales caractéristiques morphologiques de l'écriture amazighe. Dans la quatrième section nous présentons les différentes opérations de prétraitement. Dans la cinquième section, nous décrivons la méthode d'extractions des caractéristiques. Dans la sixième section nous présentons la méthode de classification.

2. Théorie des chaînes de Markov

2.1 Définition

Un processus stochastique $\{X_t, t \in T\}$est une collection de variables aléatoire définies sur un même espace probabilisé, l'indice t est souvent interprété comme le temps (Pechwitz et al, 2000). Le processus est en temps continu si T est continu, et en temps discret si T est discret.

La variable X_t représente l'état du processus au temps t et l'ensemble de toute les valeurs possibles pour cette variable est appelé l'espace des états du processus et sera noté E (Pechwitz et al, 2000).

Un processus stochastique dont l'ensemble des états E est fini ou dénombrable est appelé une chaîne (Pechwitz et al, 2000). Un processus est à temps discret lorsque l'ensemble T est fini ou dénombrable.

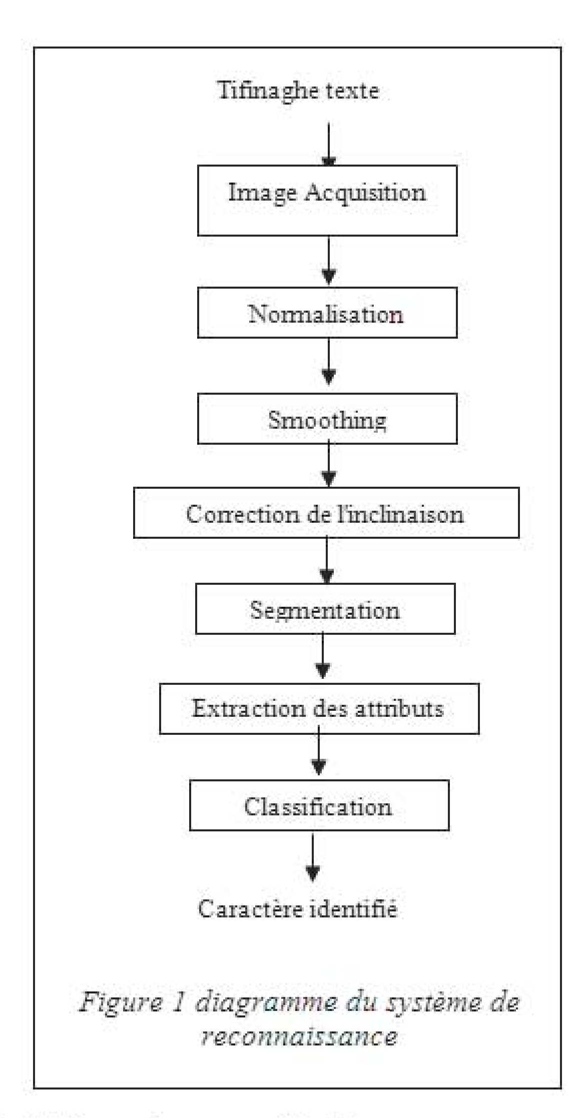


Figure 1 diagramme du système de reconnaissance

Une chaîne de Markov à temps discrète est un processus stochastique $\{X_n, n = 0,1,...\}$ à temps discret, défini sur un espace d'états E fini ou dénombrable et vérifiant la propriété de Markov

$$P[X_n = i | X_0, ... X_{n-1}] = P[X_n = i | X_{n-1}], \text{pour tout } i \in E \quad et \quad \forall n \geq 1.$$

2.2 Chaîne Observable

L'évolution du processus de Markov peut être représentée par un graphe de transitions d'états (figure 2) qui fait apparaître la structure du processus selon certaines règles.

2.2.1 Chaîne Cachée

Dans un Modèle de Markov Caché (MMC) les états $S = \{s_1, s_2, ... , s_n\}$ sont non observables; cependant ils émettent des signaux observables $O = \{o_1, o_2, ..., o_n\}$ qui sont pondérés par leur probabilité. Le modèle λ peut être représenté graphiquement (Figure 3), avec les états $S = \{s_1, s_2, ... , s_n\}$

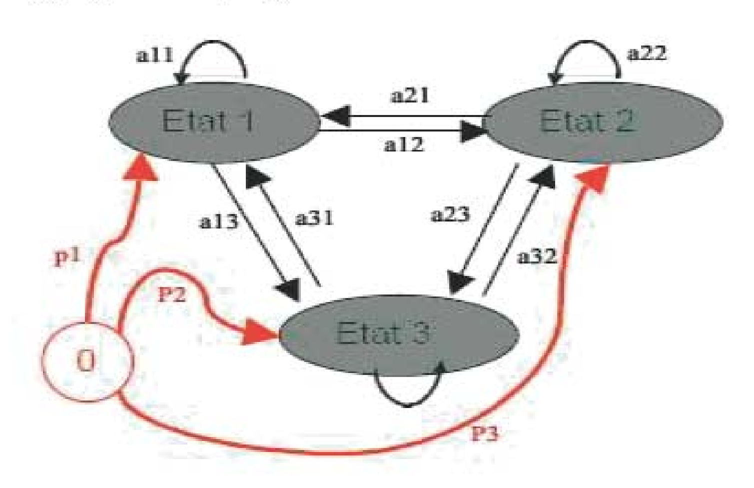


Figure 2 Graphe d'un Modèle de Markov Observable

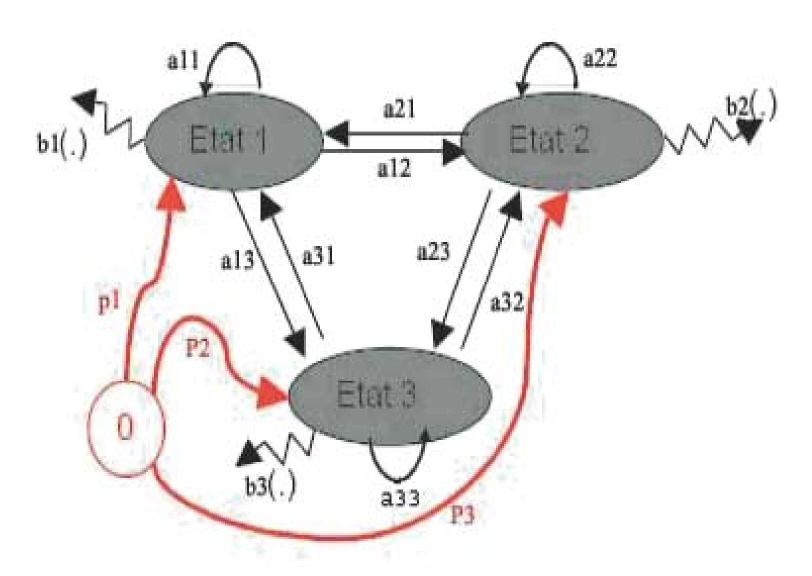


Figure 3 Graphe d'un Modèle de Markov cachée

2.2.2 *Types de HMMs*

Selon la topologie du réseau des états, il y a deux types de HMMs. Ce sont le modèle ergodique et le modèle gauche droite.

2.2.3 *Modèle ergodique*

Modèle ergodique c'est un modèle sans contraintes où toutes les transitions d'un état vers l'autre sont possibles, c'est-à-dire $a_{ij} \succ 0 \quad \forall(i,j) \in [1,N]$.

2.2.4 *Modèle gauche droite*

Modèle gauche droite (figure 3) est un modèle où il y a des contraintes sur des transitions : seulement la transition d'un état ayant un indice bas vers un état ayant un indice haut est possible.

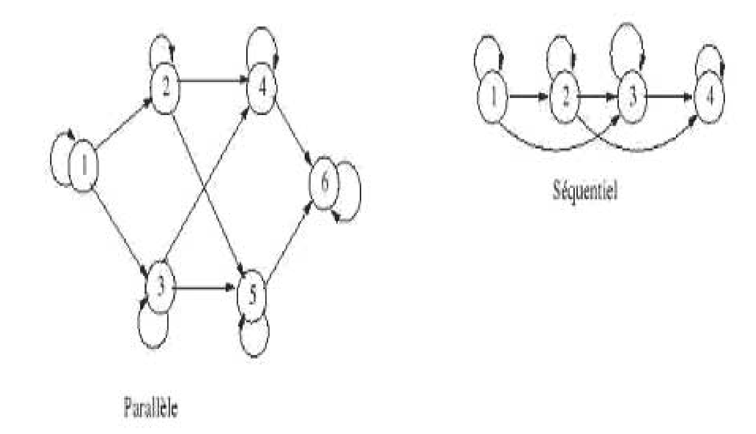


Figure 4 Les modèles gauches droits.

2.3 Problèmes pour HMM

Problème 1 : Étant donné le modèle λ = (A,B, π), comment on peut calculer P (O|λ), la probabilité de l'occurrence de la séquence des observations $O = o_1, o_2, ..., o_T$ c'est le problème de reconnaissance. Pour résoudre ce problème, on utilise la procédure aller-retour (forward-backward procédure).

Problème 2 : Étant donné le modèle λ = (A,B, π), comment on peut choisir une séquence des états $Q = q_1, q_2, ..., q_T$ afin de maximiser la probabilité P (O,Q|λ). Pour résoudre ce problème on utilise l'algorithme de Viterbi.

Problème 3 : Comment on peut ajuster les paramètres de HMM afin de maximiser P (O|λ).

C'est le problème d'apprentissage. Pour résoudre ce problème on utilise l'algorithme de Baum - Welch.

3. Les caractéristiques morphologiques des caractères amazighs.

Le Tifinaghe est le système d'écriture de la langue amazighe. Il tire son origine du vieil alphabet libyque et saharien, déjà utilisé depuis le VIème siècle avant l'ère chrétienne par les populations de l'Afrique du Nord, du Sahel et des Iles Canaries. Cet alphabet a subi des modifications et des variations depuis son origine jusqu'à nos jours.

La figure 4 ci-dessous présente les différents modèles de l'alphabet Amazigh (Tifinaghe-IRCAM). Il comporte cinquante cinq lettres. A la différence des caractères latins et arabes, l'écriture Amazighe n'est jamais cursive, ce qui facilite toute opération de segmentation. La majorité des modèles graphiques des caractères est composée de points, de petits cercles, et/ou de segments. De plus, les segments sont tous verticaux, horizontaux, ou diagonaux.

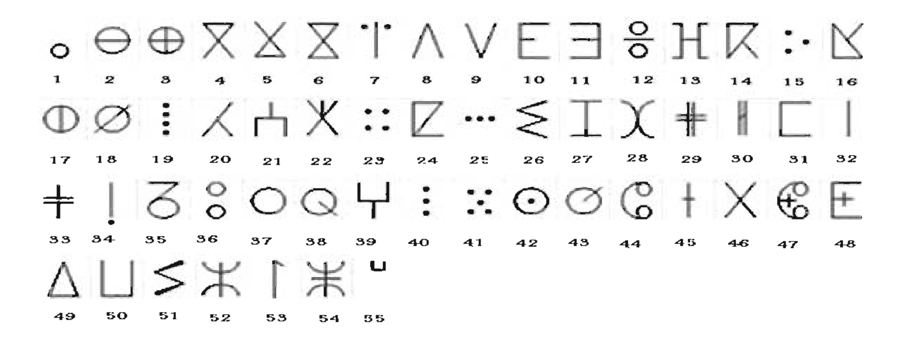


Figure 5 les caractères Amazighs

4. Prétraitement

4.1 Seuillage

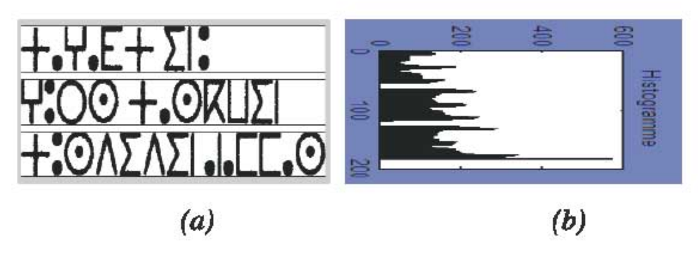
L'image entrée est une image en couleur et les algorithmes de reconnaissance courants travaillent souvent sur des images binaires. Donc, il faut faire le seuillage. Pourtant, quand le fond est très compliqué, cela devient un problème difficile.

4.2 Réduction du bruit

Le problème du bruit est très important mais très difficile à réduire en totalité. Ici nous avons adopté le filtre médian (El-Hajj, 2007).

4.3 Segmentation

Pour détecter des lignes, on peut utiliser la projection verticale comme dans l'image suivante :



(a) *(b)*

Figure 6 (a) détection des lignes (b) histogramme Horizontal

Cependant, quand les lignes sont longues et obliques, la détection des lignes devient plus difficile. De plus, les mots et les caractères dans ces lignes sont rotatoires. Cela cause des problèmes pour l'étape de reconnaissance. Donc, il faut appliquer les techniques de détection d'obliquité. Pour détecter des mots dans une ligne, on utilise la projection horizontale. La difficulté est de déterminer la distance entre les mots dans une ligne pour qu'on puisse combiner les parties isolées d'un mot (Märgner, 2005 ; Amrouch, 2009 ; El ayachi 2009).

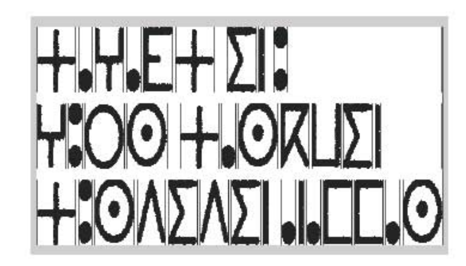


Figure 7 Résultat de segmentation vertical

Si l'étape de reconnaissance est basée sur le caractère ou sous caractères, il faut les segmenter. Pour l'écriture manuscrite c'est un problème particulièrement difficile parce qu'il n'y a pas des points pour les séparer.

4.4 Normalisation

La taille d'écriture peut varier largement. Pour faciliter l'étape de reconnaissance, il faut normaliser l'image entrée en une taille fixée. Mais si la taille fixée est très petite, on peut perdre d'information, si elle est très grande, l'étape de reconnaissance va opérer lentement.

5. Extraction des caractéristiques

L'application des MMCs à la reconnaissance de l'écriture se ramène généralement à la transformation de la forme en primitives judicieusement choisies,

L'identification directe du caractère à partir de son image (matrice de pixels) semble très difficile même impossible à cause de la morphologie des caractères amazighs et de la grande variabilité liée au style d'écriture utilisé et au bruit entachant l'image. D'où la nécessité d'obtenir, à partir de la représentation en pixels du caractère, un ensemble de primitives permettant d'identifier facilement ce dernier. Ces caractéristiques doivent être discriminantes.

Afin d'extraire ces primitives à partir de l'image du caractère, nous effectuons d'abord la Transformée de Hough (Fakir, 2009) de chaque image.

5.1. La transformée de Hough

La transformée de Hough est une méthode classique de détection de formes simples dans une image souvent utilisé pour l'extraction de primitives (Fakir, 1993). L'approche adoptée par cet algorithme est de chercher à accumuler à l'intérieur d'un espace de paramètres représentatifs, des données confirmant la présence de formes particulières.

Nous proposons dans cet article de détecter les droites à l'aide de cet algorithme :

5.2. Exemple

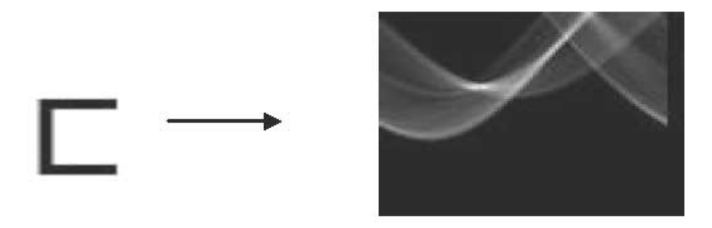


Figure 8 : La lettre M et sa transformée de Hough

Dans les calculs, les images de caractères sont de taille fixe (96*96) pixels, et le Δθ est pris égale 30, par conséquent l'accumulateur de Hough contiendra une information sur 6 orientations de (0°, 30°, 60°, 90°, 120°, 150°).

5.3. Génération des séquences d'observation

La génération de la séquence d'observations directionnelles est obtenue en exploitant les données enregistrées lors de la partie précédente. En effet, nous sélectionnons le minimum des primitives représentatives pour les directions dominantes.

La figure ci-dessous représente le vecteur d'observation généré par la lettre M.

(5, 6, 1, 4, 5, 6 ,1 ,3 ,4 ,5, 6, 1 ,2, 3)

6. Classification

Lors de cette phase nous entraînons les modèles de Markov cachés de différents caractères par la procédure classique de Baum-Welch afin d'ajuster leurs paramètres. Chaque caractère possède son propre modèle suivant les résultats de l'étape précédente. Par conséquent l'algorithme va rechercher dans tout l'espace des MMC modélisant chaque caractère, celle qui a la probabilité maximum de générer la séquence d'observations constituée à la phase « génération de la séquence d'observation » précédente. Le meilleur MMC trouvé, est enregistré pour former une base d'apprentissage

Nous avons construit durant l'apprentissage autant de MMCs qu'il y avait d'images de caractères à apprendre, alors la classification se fait d'abord par recherche du modèle discriminant parmi tous les meilleurs MMCs enregistrés pendant cette phase de l'ensemble des caractères étudiés. En effet, nous calculons par l'algorithme de Forward avec quelle probabilité ces modèles peuvent générer la séquence d'observations de caractère à reconnaître, par la suite nous disposons d'un ensemble de modèles avec chacun un score, le modèle élu est celui possédant le plus grand score.

7. Résultats expérimentaux

42 caractères sont lus dont 21 ont été reconnus, soit un taux de reconnaissance de 87%. En ce qui concerne les lettres, le meilleur résultat atteint avec cette approche a été de 94%, pour le caractère (zed).

8. Conclusion et perspectives

Les Modèles de Markov Cachés s'adaptent bien à la variation de la longueur d'écriture manuscrite, Cependant, sa capacité de discriminante n'est pas très forte car chaque MMC utilise les données d'apprentissage d'un seul caractère. De plus l'une des faiblesses des MMCs, provient du niveau de l'estimation des probabilités d'émission d'observations. Pour remédier à ces problèmes, nous pensons à une méthode hybride combinant les MMCs et les réseaux de neurones, ou à une hybridation des MMCs et SVM.

Références

Koerich A., Sabourin R. Suen S. (2003) Large vocabulary Hand-writing Recognition: A survey *Pattern Analytical Applied June* 97-121.

Olliver D., Weinfeld M. and Guegan R. (2000). Combining Dinerent Classic

El-Hajj R., Mokbel C., Likforman-Sulem L.,52007- Combination of HMM-based classifiers for the recognition of Arabic Handwritten Words, *Actes d'ICDAR'07*, pp. 959-963, Curitiba – Brazil,

Pechwitz M., Märgner V. (2003), HMM Based approach for handwritten Arabic Word Recognition Using the IFN/ENIT– DataBase, *ICDAR'03, Edinburgh*, pp. 890-894.

Märgner V., Pechwitz M., El-Abed H., (2005). Arabic Handwriting Word Recognition Competition, *Actes d'ICDAR'05*, pp. 70 - 74, Seoul,

Ben Amara N., Belaïd A., Ellouze N. 2000. Utilisation des modèles markoviens en reconnaissance de l'écriture arabe : Etat de l'art, *3ème Colloque International francophone sur l'écrit et le document (CIFED'00)*, pp181-191(Lyon, FRANCE),

Amrouch, M. Es saady Y Rachidi, . A. El Yassa M. and Mammass. D. (2009). Printed Amazigh Character Recognition by a Hybrid Approach Based on Hidden Markov Models and the Hough Transform, 978-1-4244-3757-3/09/$25.00 *©2009 IEEE*

Fakir M., Bouikhalene B. and Moro K. (2009). Skeletonization Methods Evaluation for the Recognition of PrintedTifinaghe characters, *SITCAM'09*, Agadir-Maroc

Fakir M. and Sodeyama C. (1993). Recognition of Arabic printed Scripts by Dynamic Programing Matching Method, *IECICE Trans. Inf & Syst,* Vol. E76- D, No.2 Feb. 93, pp: 31-37

Fakir M., Hassani M.M. and Sodeyama. C. (2000). On the recognition of Arabic characters using Hough transform technique, Malysian Journal of Computer Science Vol. 13, No.2, Dec.2000, pp: 39-47.

Brown. M. K. (1983). Pre-processing techniques for cursive word recognition, Pattern Recognition,Vol.13, N°.5, pp: 447-451, 1983

Hu M. K. (1962). Visual pattern recognition by moment invariants, *IRE trans. Infor. Theory TT-8,* pp: 179-187, 1962.

El ayachi R. and Fakir M. (2009). Recognition of Tifinaghe Characters Using Neural Network, 978-1-4244-3757-3/09/$25.00 *©2009 IEEE*

Es saady Y., Amrouch M., Rachidi A., El Yassa M. and Mammass D. (2009). Reconnaissance de caractères Amazighes Imprimés par le Formalisme des Automates à états finis, *SITCAM'09*, Agadir-Maroc,12-13 December 2009, pp:48-57.

Rachidi A., Mammass D.. (2005), Informatisation de La Langue Amazighe: Méthodes et Mises En OEuvre, SETIT 2005 3rd International Conference: Sciences of Electronic Technologies of Information Telecommunications March 27-31, 2005 – TUNISIA.

Amrouch M., Es Saady Y., Rachidi A., Elyassa M., Mammass D. (April 2009), Printed Amazigh Character Recognition by a Hybrid Approach Based on Hidden Markov Models and the Hough Transform, ICMCS'09, Ouarzazate-Maroc.

BOUIKHALENE B., ELKESSAB B. , FAKIR , M. SAFI S.(2009), La reconnaissance des caractères Tifinagh par l'utilisation des réseaux de neurones multicouches ICMC'9, Maroc.

Construction et exploitation d'un lexique morpho-syntaxique des verbes arabes

Abdelhamid El Jihad, Said El Hassani, Salim Rami

Institut d'Etudes et de Recherches pour l'Arabisation
{jihad.hamid, said.elhassani, salim.rami}@gmail.com

Résumé

La question des ressources lexicales est de première importance dans le domaine du Traitement Automatique des Langues. En effet, le développement d'applications efficaces reposant sur le traitement d'une langue donnée exige des ressources linguistiques de qualité. A l'exception de la langue Anglaise, il est constaté que de telles ressources sont encore à l'état embryonnaires pour des langues telles que le Français ou l'Espagnol, et pratiquement inexistantes pour l'arabe. Cette absence est la conséquence directe de deux facteurs : les coûts de constitution de telles ressources et le temps nécessaire à leur élaboration et finalisation.

Nous proposons de présenter au sein de ce travail une ressource lexicale pour l'arabe. Il s'agit d'un lexique morpho-syntaxique des verbes arabes conjugués à large couverture et utilisable dans les applications de Traitement Automatique des Langues.

La ressource contient 24175 verbes distincts, pour lesquels nous générons à partir d'un conjugueur les formes conjugués entièrement voyellés (soit environ 2446962 entrées). Des informations morpho-syntaxique (temps, voix, case, personne, nombre, genre) sont données en plus du lemme et de la racine dont découle la forme conjuguée.

1. Introduction

Les lexiques morpho-syntaxiques sont des ressources fondamentales pour le traitement automatique des langues. Ils associent un mot à une ou plusieurs catégorie/s grammaticale/s et un ou plusieurs lemme/s. Aujourd'hui, de très nombreux lexiques ont été produits dans le domaine, et sont majoritairement de

langue étrangère (anglais, français...). Ceci a permis l'essor considérable des traitements automatiques concernant ces langues.

Pour la langue arabe, il n'existe pas à ce jour de lexiques morpho-syntaxiques aisément disponibles et les recherches linguistiques qui ont recours à des lexiques morpho-syntaxiques sont rares. Soucieux plus par ce manque, l'équipe Traitement Automatique de la Langue arabe a entrepris un projet de recherche dont l'objectif est la constitution d'un lexique morpho-syntaxiques. La disponibilité de ce lexique va donner un coup d'envoi aux divers travaux de recherche linguistique qui utilisent les lexiques morpho-syntaxiques.

2. Classification des verbes arabes

2.1. Généralités

Tout verbe arabe est formé sur une racine de trois ou quatre consonnes coulée dans un ou plusieurs schèmes caractéristiques, on parle alors respectivement de verbes trilitères ou verbes quadrilitères. Selon la nature des lettres qui forment la racine on distingue deux principales classes: verbes réguliers et verbes à glides.

2.2. Verbes réguliers et verbes à glides

2.2.1. Verbes réguliers

La classe des verbes réguliers (الأفعال الصحيحة) est formée des verbes dans lesquels aucune des lettres radicales n'est faible (ي ou و), cette classe est formée de trois sous classes:

a) Les verbes à racine saine (الأفعال السالمة): ce sont des verbes dans lesquels la Hamza ne constitue pas une lettre radicale, et la deuxième et la troisième lettres radicales ne peuvent être identiques, par Exemple : (كَتَبَ).

b) Les verbes à racine redoublé (الأفعال المضاعفة): ce sont des verbes dans lesquels la deuxième lettres radicales est doublée (la deuxième et la troisième lettres radicales). Par Exemple: (عَدَّ).

c) Les verbes hamzes (الأفعال المهموزة): ce sont des verbes où la Hamza constitue l'une des lettres radicales. Par Exemple : (قَرَأَ).

2.2.2. Verbes à glides

La classe des verbes à glides(الأفعال المعتلة)) sont des verbes dans lesquels une ou deux lettres radicales est/sont faible/s (ي ou و) cette classe regroupe cinq sous classes qui sont :

a) Les verbes assimilés (الأفعال المثالية): ce sont des verbes dans lesquels la première lettre radicale est faible (و ou ي). Ils ont appelés ainsi parce qu'ils sont assimilés aux verbes sains et se conjuguent de la même manière qu'eux à l'accompli actif et passif. Par Exemple : (وَعَدَ: assimilé waw) et (يَسَرَ: assimilé yae).

b) Les verbes concaves (الأفعال الجوفاء): ce sont des verbes dans lesquels le deuxième radical est (و ou ي), il est ainsi appelé parce que la lettre faible se trouve au milieu. Exemple: (قَالَ, بَاعَ ,..).

c) Les verbes manquants ou défectueux (الأفعال الناقصة): ce sont des verbes dans lesquels la dernière lettre radicale est faible. Exemple: (دَعَـــا, رَمَى,…).

d) Les verbes dits (اللفيف المفروق) Si la première et la troisième lettre radicale sont faibles, par exemple : (وَقَى, يَدَى).

e) Les verbes dits اللفيف المقرون. Si la première et la deuxième ou la deuxième et troisième lettre radicale sont faibles, par exemple : (هَوِيَ , سَوِيَ ,…)

2.3. Verbes simples et verbes augmentés

Pour certains verbes, la forme de citation contient trois ou quatre consonnes qui forment la racine et des voyelles brèves choisi parmi (a, u, i). Ces verbes sont appelés des verbes simples.

Pour d'autres verbes, on trouve dans la forme de citation, autre les consonnes de radicales, soit des voyelles longues, soit une ou plusieurs des dix consonnes formatives de schèmes ces verbes sont appelés des verbes augmentés.

2.3.1. les schèmes du verbe simple:

Les schèmes du verbe simple sont caractérisés par une variation de la deuxième voyelle du verbe à l'accompli et à l'inaccompli (alternance vocalique). Six types d'alternances vocaliques existent:

1) فَعَلَ – يَفْعِلُ نحو: جَلَسَ – يَجْلِسُ

2) فَعَلَ – يَفْعُلُ نحو: نَصَرَ – يَنْصُرُ

3) فَعَلَ – يَفْعَلُ نحو: ذَهَبَ – يَذْهَب

4) فَعِلَ – يَفْعَلُ نحو: مَرِضَ – يَمْرَضُ

5) فَعِلَ – يَفْعِلُ نحو: حَسِبَ – يَحْسِبُ

6) فَعُلَ – يَفْعُلُ نحو: كَرُمَ – يَكْرُمُ

On constate que les schèmes du verbe simple peuvent être identifiés par la forme de l'accompli: ce sont les schèmes (فَعَلَ, فَعُلَ, فَعِلَ)

2.3.2. les schèmes du verbe augmentés:

Théoriquement, il y a quinze schèmes augmentés de verbes en arabes. En pratique, dix d’entre elles se rencontrent avec une certaine fréquence. Les autres sont des formes rares.

7) فَعَّلَ - يُفَعِّلُ نحو: صَدَّقَ – يُصَدِّقُ

8) فَاعَلَ - يُفَاعِلُ نحو: كَاتَبَ – يُكَاتِبُ

9) أَفْعَلَ - يُفْعِلُ نحو: أَكْرَمَ – يُكْرِمُ

10) تَفَعَّلَ - يَتَفَعَّلُ نحو: تَقَطَّعَ – يَتَقَطَّعُ

11) تَفَاعَلَ - يَتَفَاعَلُ نحو:تَرَاسَلَ – يَتَرَاسَلُ

12) اِنْفَعَلَ - يَنْفَعِلُ نحو: اِنْكَسَرَ – يَنْكَسِرُ

13) اِفْتَعَلَ - يَفْتَعِلُ نحو: اِخْتَبَرَ – يَخْتَبِرُ

14) اِفْعَلَّ - يَفْعَلُّ نحو: اِحْمَرَّ يَحْمَرُّ

15) اِسْتَفْعَلَ - يَسْتَفْعِلُ نحو: اِسْتَسْلَمَ – يَسْتَسْلِمُ

16) اِفْعَالَّ - يَفْعَالُّ نحو: اِحْمَارَّ – يَحْمَارُّ

17) اِفْعَوْعَلَ - يَفْعَوْعِلُ نحو: اِخْشَوْشَنَ –يَخْشَوْشِنُ

18) اِفْعَوَّلَ - يَفْعَوِّلُ نحو: اِجْلَوَّذَ – يَجْلَوِّذُ

19) فَعْلَلَ - يُفَعْلِلُ نحو: دَحْرَجَ– يُدَحْرِجُ

20) تَفَعْلَلَ - يَتَفَعْلَلُ نحو تَدَحْرَجَ– يَتَدَحْرَجُ

21) اِفْعَنْلَلَ - يَفْعَنْلِلُ نحو احْرَنْجَمَ–يَحْرَنْجِمُ

22) اِفْعَلَلَّ - يَفْعَلِلُّ نحو اقْشَعَرَّ– يَقْشَعِرُّ

Le nombre de paradigme de conjugaison peut être calculé à partir des différentes formes précités en tenant compte des variations phonologique et orthographique

induit par les glides. Ainsi, nous avons établie une liste de 236 paradigmes de conjugaison différente.

Pour chaque modèle de verbe, des règles morphologiques appropriées ont été conçues. Ces règles ont été reconstituées à partir des manuels de la morphologie de l'arabe standard.

3. Les principaux traits morphologiques du verbe

Un verbe arabe peut avoir six traits morphologiques :

3.1. L'aspect :

Le verbe arabe a deux aspects de conjugaison:

a) L'accompli qui exprime une action achevée.

b) L'inaccompli qui exprime une action qui est en train de se réaliser, sans être accompli.

A cela s'ajoute l'impératif qui exprime l'ordre ou la demande et dont la forme se construit à partir de celle de l'inaccompli apocopé.

3.2. Le mode :

La notion de mode n'existe que pour le paradigme inaccompli, ce dernier connaît trois modes:

a) L'indicatif.

b) Le subjonctif.

c) L'apocopé.

3.3. La personne

On en distingue trois :

a) Première personne.

b) Deuxième personne.

c) Troisième personne.

3.4. Le genre du verbe

Dans la langue arabe, il existe deux genres :

a) Masculin

b) Féminin

3.5. Le nombre du verbe

Un verbe arabe est pourvu de nombres suivants :

a) Le singulier

b) Le duel

c) Le pluriel

3.6. La voix

La langue arabe a deux voix :

a) L'actif

b) Le passif

En conséquence un verbe arabe peut avoir 126 formes fléchies quand il se conjugue à la fois la voix active et la voix passive et 70 formes fléchies lorsqu'il ne se conjugue qu'à la voix active.

4. Processus de la conjugaison

Tous les verbes arabes se conjuguent de manière identique, en ajoutant un ensemble défini de suffixes à une base verbale d'accompli, et un autre ensemble défini de préfixes de personnes et de suffixes de genre, de nombre et éventuellement de mode à une base verbale d'inaccompli. Cette opération entraîne, dans certains cas, des problèmes de voisinage qui nécessitent certains ajustements phonétiques puis, par conséquence, graphiques. On va se limiter dans ce qui suit à présenter quelques règles qui sont communes à tous les modèles

Si R3 = "ت" alors (سَكَتْ + تُ ← سَكَتُّ)

Si R3 = "ن" alors (سَكَنْ + نَ ← سَكَنَّ)

Si R1 = "ت" et catégorie = 13 (اِفْتَعَلَ–يَفْتَعِلُ) alors (اتْتَهَمَ ← اِتَّهَمَ)

Si R1 = "ن " et catégorie = 12 (اِنْفَعَلَ–يَنْفَعِلُ) alors (انْنَمَسَ ← اِنَّمَسَ)

Si (R1 = "ذ" ou R1 = "ز") et catégorie = 13 (اِفْتَعَلَ–يَفْتَعِلُ) alors (اِذْتَكَرَ ← اِذدَكَرَ,
اِزْتَهَرَ ← اِزْدَهَرَ)

Si (R1 = "ص" ou R1 = "ض" ou R1 = "ظ") et classe = 13 (اِفْتَعَلَ–يَفْتَعِلُ) alors (اِظْتَلَمَ ← اِظَّلَمَ ← اِضْتَرَبَ, اِضْطَرَبَ ← اِصْتَدَمَ, اِصْطَدَمَ)

Si R1 = "ث" et catégorie = 13 (اِفْتَعَلَ–يَفْتَعِلُ) alors (اِثْتَغَرَ ← اِثْثَغَرَ ← اِثَّغَرَ)

Si R1 = "د" et catégorie = 13 (اِفْتَعَلَ–يَفْتَعِلُ) alors (اِدْتَخَرَ ← اِدْدَخَرَ ← اِدَّخَرَ)

Si R1 = "ط" et catégorie = 13 (اِفْتَعَلَ–يَفْتَعِلُ) alors (اِطْتَلَعَ ← اِطْطَلَعَ ← اِطَّلَعَ)

5. Description de la base utilisée

Notre base couvre la plupart des verbes trilitères ainsi que les verbes quadrilitères rarement utilisé, elle contient 24179 verbes. Elle est au format xml:


```xml
<?xml version="1.0" encoding="windows-1256" ?>
<verbes>
<verbe valeur="بَتَعَ" Ina="يَبْتِعُ" R1="ب" R2="ت" R3="ع" R4="" categorie="1" classe_numero="1" transitivite_numero="2" />
<verbe valeur="بَتَكَ" Ina="يَبْتِكُ" R1="ب" R2="ت" R3="ك" R4="" categorie="1" classe_numero="1" transitivite_numero="2" />
.
.
.
</verbes>
```


Figure 1 : Extrait du fichier xml de la base utilisé

Chaque ligne contient en plus de la valeur du verbe, les informations suivantes:

- Le verbe conjugué à l'inaccomplie (attributs XML ina).
- La Racine (attributs XML R1, R2, R3 etR4).
- La Catégorie (attributs XML categorie). (Voir annexe1)
- La Classe (attributs XML classe_numero). (Voir annexe3)
- La Transitivité (attributs XML transitivite_numero). (Voir annexe2)

Cette ressource est en constante évolution.

6. Jeu d'étiquettes utilisé

Les étiquettes morpho-syntaxiques indiquent la nature de la forme qui est codée : temps, voix, case, personne, nombre, genre… La taille du jeu d'étiquette dépend

étroitement des objectifs de la recherche envisagée. Selon le niveau de finesse que l'on veut atteindre dans la description, on peut recourir à un jeu d'étiquette important. Inversement, dans certains cas, le jeu d'étiquette peut en rester à des distinctions relativement grossières. Notre jeu d'étiquette comporte 126 étiquettes différentes.

```
<?xml version="1.0" encoding="windows-1256" ?>
<Etiquettes>
<Etiquette Code="1" Temps="الماضي" Voix="المعلوم" Mode="---" Personne="1" Nombre="مفرد" Genre="محايد" />
<Etiquette Code="2" Temps="الماضي" Voix ="المعلوم" Mode="---" Personne="1" Nombre ="مثنى" Genre="محايد" />
<Etiquette Code="3" Temps="الماضي" Voix ="المعلوم" Mode="---" Personne="1" Nombre ="جمع" Genre="محايد" />
<Etiquette Code="4" Temps="الماضي" Voix ="المعلوم" Mode="---" Personne="2" Nombre ="مفرد" Genre="مذكر" />
<Etiquette Code="5" Temps="الماضي" Voix ="المعلوم" Mode="---" Personne="2" Nombre="مفرد" Genre="مؤنث" />
.
.
.
</Etiquettes>
```

<u>Figure 2 : jeu d'étiquette utilisé</u>

Chaque ligne contient les informations morpho-syntaxique (Temps, Voix, Mode, personne, Nombre, Genre) pour chaque étiquette.

7. Description de la ressource générée

Le lexique morpho-syntaxique produit est constitué de 2446962 entrées, et il est présenté sous deux format: format Txt et format Xml.

Pour le format Txt, le lexique est formé de sept colonnes séparées par des tabulations:

Colonne N° 1 contient la forme conjuguée.

Colonne N° 2 contient le lemme.

Colonne N° 3 contient l'étiquette associée à la forme conjuguée.

Colonne N° 4 contient la racine dont découle le verbe.

Colonne N° 5 contient la catégorie du verbe. (Voir annexe1)

Colonne N° 6 contient la classe du verbe. (Voir annexe3)

Colonne N° 7 contient la Transitivité du verbe. (Voir annexe2)

Forme	**Lemme**	**Etiquette**	**Racine**	**Catégorie**	**Classe**	**Transitivité**
حَسَبْتُ	حَسَبَ	1	حسب	5	1	2
حَسَبْنَا	حَسَبَ	2	حسب	5	1	2
حَسَبْنَا	حَسَبَ	3	حسب	5	1	2
حَسَبْتَ	حَسَبَ	4	حسب	5	1	2
حَسَبْتِ	حَسَبَ	5	حسب	5	1	2
. . .						

Figure 3 : Extrait du lexique généré (format txt)

D'autre part, et afin de faciliter les échanges de cette ressource à travers la communauté du TALN (indépendamment des plates-formes, des logiciels, des systèmes d'exploitation), nous allons la représenter au format Xml.

```
<?xml version="1.0" encoding="windows-1256" ?>

<formes>

<forme valeur="حَسِبْتُ" Lemme="حَسِبَ" Etiquette="1" Racine="حسب" categorie="5" classe_numero="1" transitivite_numero="2" />

<forme valeur="حَسِبْنَا" Lemme="حَسِبَ" Etiquette="2" Racine="حسب" categorie="5" classe_numero="1" transitivite_numero="2" />

<forme valeur="حَسِبْنَا" Lemme="حَسِبَ" Etiquette="3" Racine="حسب" categorie="5" classe_numero="1" transitivite_numero="2" />
```

```
<forme valeur="حَسِبْتَ" Lemme="حَسِبَ" Etiquette="4" Racine="حسب" categorie="5" classe_numero="1"
transitivite_numero="2" />

<forme valeur="حَسِبْتِ" Lemme="حَسِبَ" Etiquette="5" Racine="حسب" categorie="5" classe_numero="1"
transitivite_numero="2" />

.

.

.

</formes>
```

Figure 4 : Extrait du lexique généré (format xml)

Pour ce format, chaque ligne contient en plus de la forme conjuguée les informations suivantes:

- Le Lemme (attributs XML lemme).
- La Racine (attributs XML racine).
- L'Etiquette (attributs XML Etiquette).
- La Catégorie (attributs XML catégorie). (Voir annexe1)
- La Classe (attributs XML Classe). (Voir annexe3)
- La Transitivité (attributs XML Transitivité). (Voir annexe2)

8. Applications visées

Ce type de ressource a bien entendu un intérêt linguistique intrinsèque, mais constitue surtout un élément de base pour tout système de traitement automatique des langues. D'une manière générale, les lexiques annotés permettent de développer des outils de traitement informatique du langage, et plus particulièrement ils servent à:

- L'élaboration de systèmes: de nombreux systèmes de traitement de la langue fonctionnent par apprentissage à partir d'un lexique annoté.
- L'évaluation de systèmes: les lexiques de grande taille sont particulièrement utilisés pour évaluer les systèmes développés et valoriser les résultats des recherches.

9. Conclusion

Nous avons présenté la dernière version d'un lexique morpho-syntaxique des verbes arabes à large couverture, La disponibilité de ce lexique va donner le coup d'envoi aux divers travaux de recherche linguistique et plus particulièrement dans le domaine d'analyse morphologique, et l'étiquetage morpho-syntaxique.

Comme perspective de ce travail, on va se pencher sur l'élaboration d'un lexique morpho-syntaxiques des noms.

Références

Alghalayni M. (2000) *Collection des leçons Arabes*, librairie moderne.

Alkalak A. et al, (2007). Système de flexion et de conjugaison - règles de grammaire et de morphologie. Rapport technique, UNESCO, 7, place de Fontenoy 75352 Paris 07 SP France.

Ammar S. Dichy J. (2008).*les verbes arabes* Edition Hatier - Paris

Khemakhem A., Gargouri B., Abdelwahed A. et Francopoulo G. (2007). Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF – ISO 24613 *Proc. of TALN* ,Toulouse.

Kouloughli D. (1994). *Grammaire de l'arabe d'aujourd'hui*. Pocket - Langues pour tous.

Tourabi A. (2007). *Livre de conjugaison des verbes arabes*. Publication de l'Institut d'Etudes et de Recherches pour l'Arabisation.

Tourabi A., El jihad A. (2007). Conjugueur des verbes arabes, *Proc. of CITALA07 (2eme conférence)*, pp. 09-27.

Annexes

```
<?xml version="1.0" encoding="windows-1256" ?>
  <awzans>
    <wazn numero="1" wazn="فَعَلَ-يَفْعِلُ" />
    <wazn numero="2" wazn="فَعَلَ-يَفْعُلُ" />
    <wazn numero="3" wazn="فَعَلَ-يَفْعَلُ" />
    <wazn numero="4" wazn="فَعِلَ-يَفْعَلُ" />
    <wazn numero="5" wazn="فَعِلَ-يَفْعِلُ" />
    <wazn numero="6" wazn="فَعُلَ- يَفْعُلُ" />
    <wazn numero="7" wazn="فَعَّلَ-يُفَعِّلُ" />
    <wazn numero="8" wazn="فَاعَلَ-يُفَاعِلُ" />
    <wazn numero="9" wazn="أَفْعَلَ-يُفْعِلُ" />
```

```
    <wazn numero="10" wazn="تَفَعَّلَ-يَتَفَعَّلُ" />
    <wazn numero="11" wazn="تَفَاعَلَ-يَتَفَاعَلُ" />
    <wazn numero="12" wazn="انْفَعَلَ-يَنْفَعِلُ" />
    <wazn numero="13" wazn="افْتَعَلَ-يَفْتَعِلُ" />
    <wazn numero="14" wazn="افْعَلَّ-يَفْعَلُّ" />
    <wazn numero="15" wazn="اسْتَفْعَلَ-يَسْتَفْعِلُ" />
    <wazn numero="16" wazn="افْعَالَّ-يَفْعَالُّ" />
    <wazn numero="17" wazn="افْعَوْعَلَ-يَفْعَوْعِلُ" />
    <wazn numero="18" wazn="افْعَوَّلَ-يَفْعَوِّلُ" />
    <wazn numero="19" wazn="فَعْلَلَ-يُفَعْلِلُ" />
    <wazn numero="20" wazn="تَفَعْلَلَ-يَتَفَعْلَلُ" />
    <wazn numero="21" wazn="افْعَنْلَلَ-يَفْعَنْلِلُ" />
    <wazn numero="22" wazn="افْعَلَلَّ-يَفْعَلِلُّ" />
  </awzans>
```

<u>Annexe1 : fichier des schèmes d'un verbe arabe</u>

```
<?xml version="1.0" encoding="windows-1256" ?>
  <transitivites>
    <transitivite numero="1" transitivite="لازم" />
    <transitivite numero="2" transitivite="متعد" />
    <transitivite numero="3" transitivite="لازم أو متعد" />
  </transitivites>
```

<u>Annexe2 : fichier de transitivité d'un verbe arabe</u>

```
<?xml version="1.0" encoding="windows-1256" ?>
  <classes>
    <classe numero="1" classe="سالم" />
    <classe numero="2" classe="مهموز الفاء" />
    <classe numero="3" classe="مهموز العين" />
    <classe numero="4" classe="مهموز اللام" />
    <classe numero="5" classe="مهموز الفاء مهموز اللام" />
    <classe numero="6" classe="مضعف" />
    <classe numero="7" classe="مضعف مهموز الفاء" />
    <classe numero="8" classe="مثال واوي" />
    <classe numero="9" classe="مثال واوي مهموز العين" />
    <classe numero="10" classe="مثال واوي مهموز اللام" />
    <classe numero="11" classe="مثال واوي مضعّف" />
```

```
  <classe numero="12" classe="مثال يائي" />
  <classe numero="13" classe="متال يائي مهموز العين" />
  <classe numero="14" classe="مثال يائي مضعّف" />
  <classe numero="15" classe="أجوف واوي" />
  <classe numero="16" classe="أجوف واوي مهموز الفاء" />
  <classe numero="17" classe="أجوف واوي مهموز اللام" />
  <classe numero="18" classe="أجوف يائي" />
  <classe numero="19" classe="أجوف يائي مهموز الفاء" />
  <classe numero="20" classe="أجوف يائي مهموز اللام" />
  <classe numero="21" classe="ناقص واوي" />
  <classe numero="22" classe="ناقص واوي مهموز الفاء" />
  <classe numero="23" classe="ناقص واوي مهموز العين" />
  <classe numero="24" classe="ناقص يائي" />
  <classe numero="25" classe="ناقص يائي مهمور الفاء" />
  <classe numero="26" classe="ناقص يائي مهموز العن" />
  <classe numero="27" classe="لفيف مفروق" />
  <classe numero="28" classe="لفيف مفروق مهموز العين" />
  <classe numero="29" classe="لفيف مقرون" />
  <classe numero="30" classe="لفيف مقرون مهموز الفاء" />
</classes>
```


<u>Annexe3 : fichier des classe d'un verbe arabe</u>

ساهمت المعيرة التكنولوجية للغة الأمازيغية واستخدام حرف تيفيناغ الذي عرف دفعة قوية منذ سنة 2002 بالمغرب على وجه الخصوص، في استحداث المعالجة الآلية للغة الأمازيغية وبناء الموارد اللغوية الإلكترونية الأمازيغية بحرف تيفيناغ. غير أنه لتعزيز إدماج الأمازيغية في التكنولوجيات الحديثة والإسهام في تنميتها و تطويرها، ينبغي بذل مزيدا من الجهود، خاصة من قبل الأوساط العلمية. وتندرج الأعمال المقدمة في هذا الكتاب في هذا الإطار، حيث تصب جميعها في مجال المعالجة الآلية للغة وتتمحور حول مواضيع متنوعة. منها على سبيل المثال لا الحصر : إنشاء واستعمال القواميس الإلكترونية، تطبيقات الصرف وبناء الجملة ودلالات الألفاظ، الوسم والتحليل الصرفي والنحوي والدلالي، التعرف الضوئي على حروف تيفيناغ وغيرها.

Le processus de standardisation et de généralisation de l'utilisation du caractère tifinaghe au Maroc, promu depuis 2002 par l'IRCAM, a considérablement contribué au développement du traitement automatique de la langue amazighe et à la création des ressources linguistiques électroniques en caractères tifinaghes. Il reste néanmoins d'autres efforts à consentir, notamment par la communauté scientifique, pour encourager la promotion de l'amazighe et contribuer à son épanouissement. Les travaux inclus dans cet ouvrage y contribuent dans des axes aussi diversifiés que la dictionnairique, la morpho-syntaxe et la sémantique, l'annotation morphosyntaxique, l'OCR pour l'amazighe, etc.

30 DH